

Análise e Modelação de Preços de Anúncios Airbnb em Lisboa

ISCTE BUSINESS SCHOOL

Curso 0329: Mestrado em Ciência de Dados

Unidade Curricular: Modelos de Previsão

Grupo 5

André Colaço – nº134432
Daniel Oliveira – nº137564
Eva Jesus – nº133968
Francisco Rodrigues – nº134018
Rafael Machado – nº87341

Índice

Introdução.....	3
Business Understanding	3
Objetivo do trabalho.....	3
Preparação dos dados	3
Seleção de variáveis relevantes	3
Limpeza dos dados e tratamento de omissos.....	5
Análise e exploração dos dados.....	5
Sumário dos dados.....	5
Tratamento de outliers.....	6
Análise exploratória e descritiva das variáveis.....	7
Correlação entre variáveis	9
Divisão do Dataset.....	10
Modelos Preditivos.....	11
Modelo de Regressão Linear Múltipla	11
Avaliação de Performance do Modelo	11
Modelo de Árvore de Decisão (CART).....	13
Visualização gráfica do modelo	13
Avaliação de Performance do Modelo	14
Análise da complexidade do modelo e modelo podado	15
Modelo Bagging.....	15
Visualização gráfica do modelo	16
Avaliação de Performance do Modelo	16
Importância das variáveis e comparação com o modelo CART.....	17
Modelo Random Forest.....	17
Principais parâmetros do modelo	18
Modelo Random Forest (sem Ajuste Manual)	18
Avaliação de Performance do Modelo	18
Modelo Random Forest com Ajuste dos Hiperparâmetros	19
Resultados do Modelo Ajustado.....	20
Modelo Gradient Boosting Machine (GBM)	20
Avaliação da performance do modelo.....	21
Modelo GBM com ajuste de parâmetros.....	21
Conclusão	22

Introdução

No âmbito da disciplina de Modelos de Previsão foi-nos proposto realizar um modelo de previsão para estimar o preço de anúncios de quartos para alugar, colocados na plataforma Airbnb em Lisboa, tendo como base a aplicação dos diferentes modelos de previsão explorados em aula.

De modo a desenvolver os modelos, seguimos uma abordagem estruturada, começando pela familiarização dos dados e consequente análise e limpeza dos dados, identificando valores omissos e inconsistências, bem como explorando padrões iniciais nos preços e características dos anúncios. Posteriormente, selecionámos variáveis relevantes que poderiam influenciar o preço, como localização, tipo de quarto, avaliação dos hóspedes e tipologia do alojamento.

Após esta fase inicial de preparação, foram aplicados os modelos de previsão, seguindo-se a interpretação dos resultados obtidos, consequentemente comparámos os modelos e potenciais pontos fortes e fracos das diferentes soluções, com o objetivo de indicar o modelo com melhor performance para o problema em questão.

Business Understanding

O alojamento local tem como principal função a prestação de alojamento temporário, sendo muitas vezes uma alternativa ao setor da hotelaria tradicional. Nos últimos anos, o crescimento de plataformas digitais veio alterar significativamente este mercado, destacando-se o Airbnb como uma das plataformas mais conhecidas, permitindo “democratizar” o mercado de alojamento.

Na cidade de Lisboa, o Airbnb assume um papel relevante na oferta de alojamento turístico, contribuindo para a diversificação da oferta, mas também levantando desafios associados à regulação, à pressão sobre o mercado habitacional e à distribuição espacial das atividades turísticas.

Objetivo do trabalho

Neste contexto, o objetivo principal deste trabalho passa por analisar e caracterizar as listagens do Airbnb em Lisboa. De forma a cumprir com este objetivo, a análise deste trabalho procura responder às seguintes questões:

- Existem diferentes perfis de acomodações anunciadas no Airbnb em Lisboa?
- De que forma os preços das acomodações se relacionam com as suas características?

Preparação dos dados

Seleção de variáveis relevantes

De forma a começar a análise dos dados começamos pela importação do dataset `listings_Lisbon.csv` para o RStudio. O ficheiro encontrava-se em formato CSV, no entanto, devido à existência de algumas linhas que não seguiam o formato standard, o mesmo foi importado seguindo a estrutura de um ficheiro em texto bruto. Após a limpeza das observações com formatação incorreta, a importação, via CSV resultou num DataFrame com 18 variáveis e 25 540 observações (vs as originais 25 544 linhas).

Após importados os dados, avançámos para a avaliação da relevância de cada variável para o modelo preditivo. Abaixo encontra-se a listagem das variáveis originais presentes no *dataset*, bem como a avaliação da sua inclusão no modelo preditivo:

- Price – relevante: Preço por noite. Variável a prever com os modelos de previsão aplicados.
- ID – não relevante: Identificador único do registo, não acrescentando informação relevante ao modelo preditivo.
- Name – não relevante: Nome da listagem. Apesar de potencialmente relevante como característica preditiva no âmbito de *text mining*, a mesma não foi incluída no âmbito desta análise.
- Host_ID – potencialmente relevante: Identificador único do anfitrião. Pode existir uma relação entre o anfitrião e o preço dos seus alojamentos, pelo que tentámos manter este valor inicialmente (no entanto devido à existência de um elevado número de níveis, não foi considerado na criação dos modelos).
- Host_name – não relevante: Nome do anfitrião. Não contribui para o modelo, visto mantermos já o id.
- Neighbourhood_group – relevante: Freguesia da localização do alojamento. Relevante visto ser uma das unidades mais pequenas identificadoras de localização, e captura de forma mais relevante tendências de localizações onde o alojamento local é mais comum (como a baixa lisboeta).
- Neighbourhood – relevante: Concelho da localização do alojamento. Relevante visto ser mais uma unidade relevante de localização, apesar de menos específica.
- Latitude – não relevante: Indicador potencialmente útil para clustering, mas como não aplicámos técnicas não supervisionadas, foi excluída.
- Longitude – não relevante: Indicador potencialmente útil para clustering, mas como não aplicámos técnicas não supervisionadas, foi excluída.
- Room_type – relevante: Tipo de alojamento associado à listagem (quarto privado/partilhado, hotel ou apartamento/casa inteira), representando potencialmente um dos mais relevantes indicadores de preço, devido à relação com a dimensão do espaço/potencial nº máximo de hóspedes.
- Minimum_nights – relevante: Número de noites mínimas de estadia. Apesar de não ser vista uma relação clara com preço, decidimos explorar na aplicação dos modelos.
- Number_of_reviews – relevante: Número de avaliações da listagem. Potencialmente relevante, visto ser indicador da qualidade do alojamento por representar uma aproximação do nº de hóspedes acomodados ao longo dos anos.
- Last_review – relevante: Data da última avaliação. Pode ser relevante, no entanto foi adaptada para apresentarmos o nº de dias desde a última avaliação.
- Reviews_per_month – relevante: Média de avaliações por mês. Pode estar relacionada com a dinâmica da procura da listagem e, por isso, com o preço.
- Calculated_host_listings_count – relevante: Nº de propriedades do anfitrião. Pode ser relevante para previsão de preço (ex: empresas com elevado nº de listagens com preço standardizado).
- Availability_365 – relevante: Nº de dias disponíveis para alugar por ano. Pode ser relevante devido à sua influência na oferta da listagem em particular e potencial “exclusividade”.
- Number_of_reviews_ltm – relevante: Nº de avaliações no último mês. Potencialmente representativo da procura atual do alojamento.
- License – não relevante: Licença associado ao alojamento. Considerada não relevante para o modelo preditivo, visto ser uma característica apenas para efeitos legais.

Após a seleção das variáveis relevantes, procedemos com as seguintes 12 variáveis: *host_id*, *neighbourhood_group*, *neighbourhood*, *room_type*, *price*, *minimum_nights*, *number_of_reviews*, *last_review*, *reviews_per_month*, *calculated_host_listings_count*, *availability_365*, *number_of_reviews_ltm*.

Limpeza dos dados e tratamento de omissos

Após a seleção das variáveis relevantes, procedemos à limpeza dos dados, uma vez que algumas variáveis apresentavam valores omissos ou problemas de formatação. As transformações aplicadas foram as seguintes:

- *Price*: Sendo a variável alvo do modelo, todas as observações com valores NA em *Price* foram removidas (3746 observações, passando para 21794).
- *Reviews_per_month*: Existiam 2417 observações com valores NA. Dado que estes valores não podem permanecer omissos, substituímos os NA por 0, visto que estes casos ocorriam apenas para as observações com *number_of_reviews* = 0.

```
> Dados2_rpm_na <- Dados2[(is.na(Dados2$reviews_per_month) | Dados2$rev == ""), ]
> View(Dados2_rpm_na)
> unique(Dados2_rpm_na$number_of_reviews)
[1] 0
>
```

Figura 1 - casos com *reviews_per_month* NA têm sempre *number_of_reviews* a 0

- *Last_review*: Visto esta variável se tratar de uma data, começamos por realizar a transformação dos dados em formato data. De modo a facilitar a utilização da informação correspondente à variável, foi criada em seguida a variável *days_since_last_review*, com o objetivo de registar o número de dias desde a última avaliação, tornando assim a variável numérica. Nos casos em que não existiam avaliações o *days_since_last_review* ficou registado como 0. A variável original *last_review* foi depois removida.

Análise e exploração dos dados

Sumário dos dados

Após o tratamento dos dados omissos, conseguimos realizar as funções `summary(Dados2)` e `str(Dados2)`, que nos dão um breve *overview* das colunas sujeitas à análise e do tipo de variáveis que representam.

```

> # -----
> # ANÁLISE EXPLORATÓRIA/IDENTIFICAÇÃO DE OUTLIERS
> # -----
> summary(Dados2)
  host_id      neighbourhood_group      neighbourhood      room_type
447375630: 339 Lisboa :14976 Santa Maria Maior: 3225 Entire home/apt:16897
3953109 : 216 Cascais: 2108 Misericrdia : 2422 Hotel room : 130
76223539 : 176 Sintra : 1386 Arroios : 1701 Private room : 4633
419162816: 161 Mafra : 1147 Cascais e Estoril: 1431 Shared room : 134
505424337: 136 Lourinh: 486 Santo Antnio : 1322
508522478: 125 Oeiras : 426 So Vicente : 1141
(Other) :20641 (Other): 1265 (Other) :10552
 price      minimum_nights      number_of_reviews      reviews_per_month
Min. : 9.0 Min. : 1.000 Min. : 0.00 Min. : 0.000
1st Qu.: 75.0 1st Qu.: 1.000 1st Qu.: 4.00 1st Qu.: 0.280
Median : 110.0 Median : 2.000 Median : 27.00 Median : 1.020
Mean : 248.4 Mean : 4.339 Mean : 76.52 Mean : 1.505
3rd Qu.: 167.0 3rd Qu.: 3.000 3rd Qu.: 98.00 3rd Qu.: 2.300
Max. : 86733.0 Max. : 730.000 Max. : 1787.00 Max. : 29.720

calculated_host_listings_count      availability_365      number_of_reviews_ltm
Min. : 1.00 Min. : 0.0 Min. : 0.00
1st Qu.: 1.00 1st Qu.:159.0 1st Qu.: 1.00
Median : 5.00 Median :285.5 Median : 8.00
Mean : 24.26 Mean :243.0 Mean : 15.35
3rd Qu.: 15.00 3rd Qu.:335.0 3rd Qu.: 24.00
Max. : 391.00 Max. : 365.0 Max. : 329.00

days_since_last_review
Min. : 0.0
1st Qu.: 90.0
Median : 102.0
Mean : 195.4
3rd Qu.: 135.0
Max. : 5258.0

```

Figura 2 - Outputs da função `summary(Dados2)`

Tratamento de outliers

Após a limpeza do *dataset*, concentrámo-nos na identificação de *outliers* e na forma mais adequada de os limitar. Ao analisar a variável *Price* verificámos que existiam numerosas observações com valores na ordem dos milhares de euros por noite, o que indicava a presença de possíveis outliers. Como observado na função *summary* os dados originais apresentavam um valor máximo de preço de 86733€, vs 3º quartil de 167€, levando a um preço médio de 248€, bastante desajustado da mediana de 110€.

Para confirmar esta suspeita, realizamos um *boxplot*, que evidenciou claramente a existência destes valores extremos e validou o nosso raciocínio inicial. O mesmo encontra-se em escala logarítmica, exatamente devido à elevada presença de outliers.

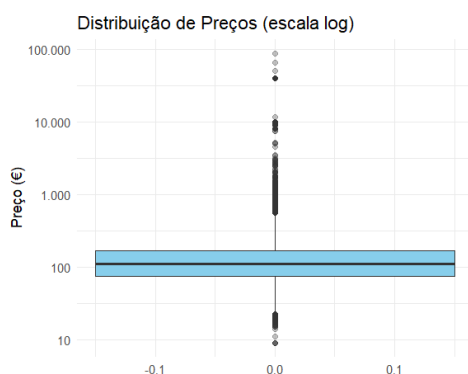


Figura 3 - Boxplot de preços

Realizámos ainda uma análise de assimetria e curtose, com a variável preço a apresentar uma assimetria muito forte ($\text{skewness} = 23,8$) e caudas extremamente pesadas ($\text{kurtosis} = 673$), refletindo a presença de muitos preços baixos/médios e alguns valores extremos muito altos. Estes indicadores reforçam a necessidade de utilizar a transformação logarítmica nos gráficos, bem como de considerar a mediana como medida central mais representativa.

Dada a informação atualmente presente no nosso dataset e a impossibilidade de avaliar algumas das condições mais relevantes para a definição do preço (como nº de quartos, casas de banho – privativa/partilhada, nº de hóspedes máximo, dimensão do espaço, segmento – luxo, económico...), decidimos optar pela remoção dos principais *outliers*. Desta forma, pretendemos focar o modelo de previsão nas listagens com preços típicos e reduzir a influência de valores extremos.

Decidiu-se assim remover os *outliers* superiores utilizando o critério do Intervalo Interquartil, aplicando um corte nos outliers em $1,5 \times \text{IQR}$ e ajustando o limite inferior para 0, dado que não existem preços negativos. Esta abordagem permite obter uma amostra mais representativa para previsão, sem distorcer os parâmetros centrais da distribuição. Com este corte, passamos para 20213 observações (vs as anteriores 21794), passando o preço máximo para 305€ e a média 117€.

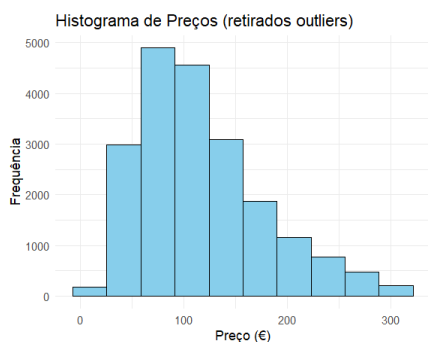


Figura 4 - Histograma de preços (retirados outliers)

Análise exploratória e descritiva das variáveis

Após o tratamento de outliers foi realizada a análise da distribuição de preços pelas variáveis categóricas mais relevantes: *neighbourhood*, *neighbourhood group* e *room_type*.

Relativamente ao *neighbourhood_group*, indicador do concelho do anúncio, vemos uma disparidade do preço médio dos anúncios, apesar de termos uma clara predominância dos anúncios colocados no concelho de Lisboa (representando 14k (70%) das 20k observações totais).

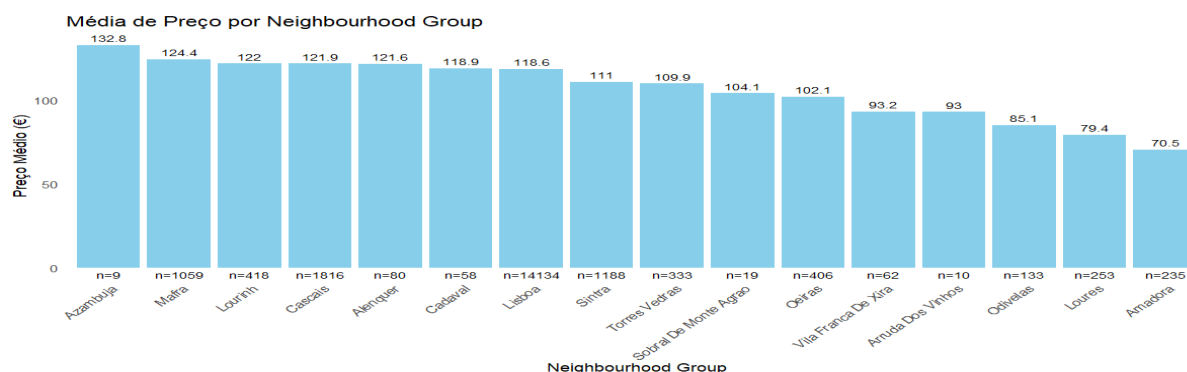


Figura 5 - Preço médio por neighbourhood group (concelho)

Relativamente ao *neighbourhood*, indicador da freguesia, temos uma maior disparidade de preços, com as 5 freguesias com preços médios mais altos a representar preços entre 3 a 4 vezes superiores às 5 freguesias com preços médios mais baixos. Apesar disso, estas freguesias

representam um número de anúncios bastante baixo (entre 31 e 2 observações).

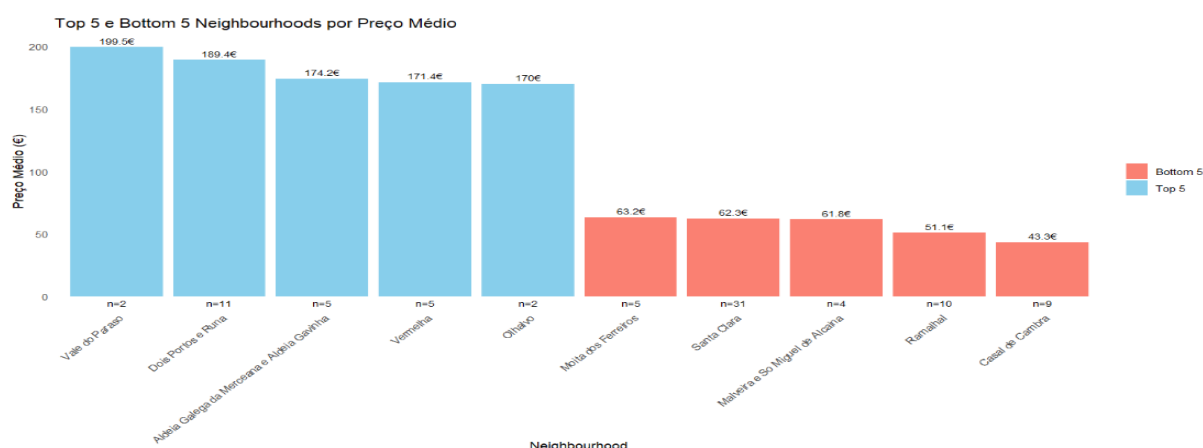


Figura 6 - Top e bottom 5 preço médio por neighbourhood (freguesia)

Outra das variáveis potencialmente mais relevantes para previsão de preço é o tipo de quarto (*room_type*). Como é possível ver no gráfico de barras abaixo, existe uma clara diferença nos preços praticados, com os quartos de hotel e alojamentos inteiros a praticarem preços mais semelhantes, na casa dos 130€ médios, contrariamente aos quartos individuais/quartos partilhados, ambos abaixo dos 70€. Apesar disso, a maioria dos anúncios representam alojamentos inteiros (~76%).

Com a realização de um *violin plot* conseguimos ainda observar que a densidade da distribuição de preços nas 4 categorias é bastante diferente, com os preços de quartos partilhados e privados a encontrarem-se mais concentrados, comparativamente às restantes categorias de alojamento.

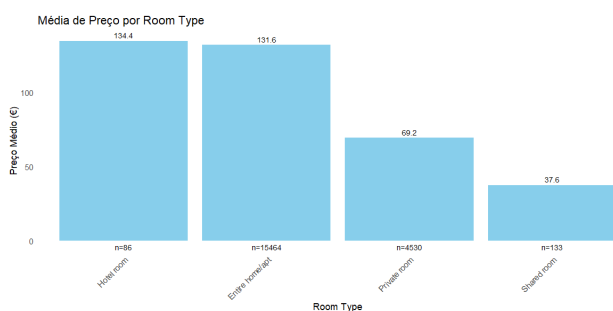


Figura 7 - Preço médio por room type

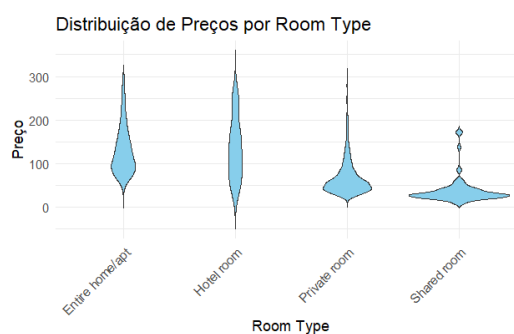


Figura 8 - Violin plot preço por room type

A distribuição das tipologias de alojamento é também diferente nos municípios do dataset, apesar de existir ainda assim uma predominância dos anúncios relativamente a alojamentos inteiros.

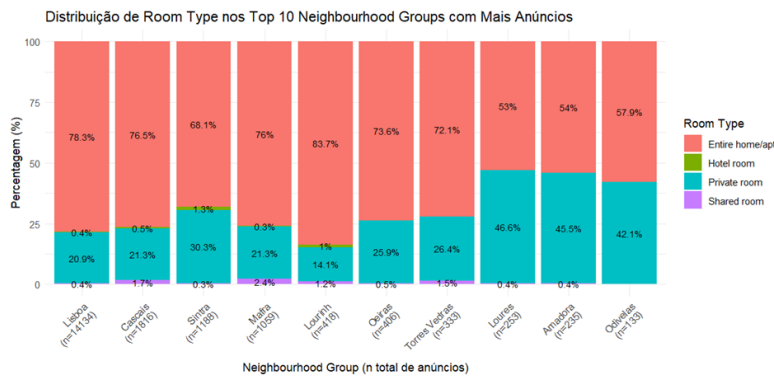


Figura 9 - Distribuição de room type por neighbourhood group (concelho)

Relativamente às restantes variáveis, utilizámos a função *describe* para nos dar um breve resumo sobre as mesmas:

```
> describe(select(Dados2_sem_outliers, 5:12), IQR = TRUE)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	IQR
price	1	20213	116.98	60.35	105.00	110.93	54.86	9	305.00	296.00	0.87	0.33	0.42	77.00
minimum_nights	2	20213	4.34	11.92	2.00	2.21	1.48	1	730.00	729.00	20.72	870.13	0.08	2.00
number_of_reviews	3	20213	78.51	117.21	29.00	52.45	41.51	0	1787.00	1787.00	2.64	10.98	0.82	95.00
reviews_per_month	4	20213	1.55	1.60	1.08	1.32	1.35	0	29.72	29.72	2.29	15.52	0.01	2.05
calculated_host_listings_count	5	20213	24.99	64.07	5.00	8.66	5.93	1	391.00	390.00	4.18	18.30	0.45	14.00
availability_365	6	20213	241.78	111.70	283.00	254.21	91.92	0	365.00	365.00	-0.78	-0.74	0.79	178.00
number_of_reviews_ltm	7	20213	15.78	19.49	8.00	12.32	11.86	0	329.00	329.00	2.33	13.29	0.14	24.00
days_since_last_review	8	20213	193.35	332.51	102.00	121.56	22.24	0	5258.00	5258.00	5.18	35.10	2.34	45.00

Figura 10 - Estatísticas descritivas das variáveis quantitativas

Relativamente à estadia mínima e disponibilidade, a variável *minimum_nights* apresenta uma forte assimetria positiva, evidenciada pela discrepância entre a média (4,34 noites) e a mediana (2 noites). Embora o intervalo interquartil indique que a vasta maioria dos anúncios exige apenas 1 ou 2 noites, existem outliers significativos com exigências de até 730 noites. Em contrapartida, a *availability_365* demonstra uma distribuição uma mediana elevada de 283 dias, sugerindo que a maior parte das propriedades mantém uma disponibilidade elevada ao longo do ano.

Por outro lado, os indicadores de avaliações (*number_of_reviews*, *reviews_per_month* e *number_of_reviews_ltm*) exibem consistentemente uma assimetria positiva. A mediana de avaliações totais é significativamente inferior à média, sendo um padrão que se repete na atividade mensal. Estes dados indicam que, embora a maioria dos anúncios receba um fluxo moderado de 1 a 2 avaliações por mês, existe um subgrupo de propriedades extremamente populares que inflaciona as médias.

Finalmente, a variável *calculated_host_listings_count* revela uma heterogeneidade na gestão das propriedades, com uma diferença substancial entre a mediana de 5 anúncios e a média de 24,99. Este desvio, aliado a um valor máximo de 391 anúncios por anfitrião representa a existência de anfitriões individuais com poucas propriedades e gestores profissionais/empresas com grandes volumes de propriedades.

Correlação entre variáveis

De forma a analisar a correlação entre variáveis começamos por analisar a relação das variáveis categóricas *room_type*, *neighbourhood* e *neighbourhood_group* com a variável preço, através do cálculo do coeficiente de Eta. Estes valores, foram, respetivamente, 0,44, 0,31, e 0,13. Estes dados indicam assim o *room_type* como a variável categórica com a associação mais forte com

o preço entre as três variáveis analisadas. Em termos de tamanho do efeito, é considerada uma associação moderada a forte.

Relativamente ao *neighbourhood* e *neighbourhood_group*, este coeficiente de correlação indica-nos assim que a localização macro (concelho – *neighbourhood_group*) não diferencia tanto o preço quanto a localização micro (freguesia - *neighbourhood*).

Para as restantes variáveis, dado serem variáveis quantitativas, foi realizada a matriz de correlação de Pearson.

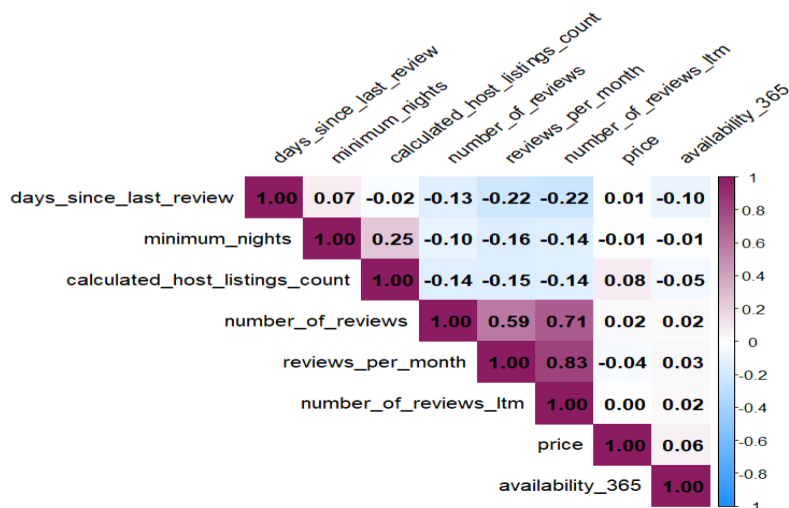


Figura 11 - Matriz de correlação de Pearson

A análise da matriz de correlação de Pearson revelou assim que as variáveis quantitativas analisadas apresentam uma correlação linear residual com a variável alvo *price* (coeficientes entre -0,04 e 0,08). Isto sugere que o preço é determinado primordialmente pelas características categóricas (como a localização e o tipo de quarto, conforme verificado anteriormente) ou por relações não-lineares, e não pelas métricas de fluxo de hóspedes ou disponibilidade. Adicionalmente, detetou-se forte correlação entre as variáveis de frequência de avaliações (entre *reviews* mensais e anuais), indicando redundância informativa.

Divisão do Dataset

A fase após a limpeza, análise exploratória e remoção de *outliers*, é a divisão do dataset em conjuntos de treino e de teste, para avaliar o desempenho preditivo dos modelos em dados.

Antes desta divisão a coluna *host_id* foi removida, pois apesar de ser relevante para identificar os alojamentos associados a cada anfitrião, apresenta um número muito elevado de níveis. Esta complexidade torna a variável pouco informativa para efeitos de previsão e suscetível de introduzir ruído no processo de modelação.

Para garantir que todas as zonas de Lisboa são representadas de forma equilibrada tanto no conjunto de treino como no conjunto de teste foi realizada uma amostragem estratificada com base na variável *neighbourhood*.

Por último, foi definida uma semente aleatória de 3.900, que permitiu replicar os resultados e evitar variações decorrentes do processo de amostragem aleatória. O dataset foi assim dividido em dois conjuntos: 70% para treino e 30% para teste, sendo esta a divisão seguida para todos

os modelos testados. O primeiro foi utilizado para ajustar os modelos preditivos e o conjunto de teste para avaliar o desempenho dos modelos.

Modelos Preditivos

Nesta fase do trabalho foram aplicados vários modelos de aprendizagem supervisionada para prever o preço de anúncios no *Airbnb* em Lisboa.

O preço é uma variável quantitativa contínua, pelo que apenas foram considerados modelos adequados a este tipo de variável. Por exemplo a regressão logística não foi utilizada, pois só se aplica quando a variável a prever é categórica.

Modelo de Regressão Linear Múltipla

A Regressão Linear Múltipla permite analisar a relação entre uma variável dependente (*price*) e várias de variáveis explicativas, tais como as características do anúncio, do alojamento e a localização.

Depois de criado, o modelo foi validado com o conjunto de teste para se poder avaliar o desempenho dos novos dados. As previsões foram obtidas com a função *predict()* e comparadas com os valores reais.

Foram também analisados os resíduos, isto é, a diferença entre os valores reais e os previstos, para verificar se o modelo cumpre, de forma geral, os pressupostos da regressão linear.

Avaliação de Performance do Modelo

Conjunto de Treino

No conjunto de treino, o modelo apresenta um RMSE de 51,89€, ou seja, em média as previsões do modelo diferem em cerca de 52€. O MAE de 39,38€, indica que a maioria dos erros de previsão é moderada, apesar de existirem alguns desvios mais elevados.

O coeficiente de determinação ($R^2 = 0,2712$; R^2 ajustado = 0,2643) mostra que o modelo explica cerca de um quarto da variabilidade do preço dos quartos.

Ao analisar o sumário percebemos ainda que o modelo é globalmente significativo, de acordo com o teste F ($F = 39,37$; $p\text{-value} < 0,001$). No entanto, a precisão das previsões é limitada pois tem um erro padrão dos resíduos de aproximadamente 52€.

Conjunto de Teste

No conjunto de teste, o modelo apresenta um RMSE de 51,16€ euros, um MAE de 38,84€ e um R^2 de 0,2718, indicando erros de previsão moderados e uma capacidade explicativa razoável. A proximidade destes valores aos obtidos no conjunto de treino sugere que o modelo generaliza bem e não apresenta problemas de overfitting.

No entanto, os testes de diagnóstico revelam limitações importantes. O teste de Breusch-Pagan indica a presença de heterocedasticidade, ou seja, os erros do modelo não têm variância constante ($BP = 750,03$; $p\text{-value} < 0,001$). Este resultado é reforçado pelo teste de White, que igualmente aponta para heterocedasticidade (estatística = 773; $p\text{-value} < 0,001$).

Adicionalmente, o teste de Breusch-Godfrey confirma a existência de autocorrelação dos

resíduos (LM = 97,192; p-value < 0,001). Estes resultados mostram que, apesar de informativo, o modelo de regressão linear múltipla não cumpre com os pressupostos fundamentais.

Análise de resíduos

A análise dos resíduos mostra que existem erros de previsão relevantes, em alguns casos o desvio ultrapassa os 50 euros, o que indica uma sobrestimação do preço real.

```
> # Cálculo dos Resíduos
> residuos<-resid(modelo_reg_lin_mult)
> # Apresentação dos 6 primeiros resíduos
> head(residuos)
      1      2      3      4      7      9
-26.23594 -49.37429 -55.80917 -29.11883 -54.62259 -27.33358
```

Figura 12 - Resíduos Modelo Regressão Linear Múltipla

Para além disso, o modelo prevê valores negativos, com um mínimo de -70,84€, o que não corresponde à realidade dos preços praticados nos anúncios do Airbnb.

Para ultrapassar esta limitação, optou-se por testar um novo modelo, transformando o preço através do logaritmo. Desta forma, as previsões passam a assumir apenas valores positivos quando regressam à escala original. Além disso, esta transformação ajuda a reduzir o impacto de valores muito elevados e contribui para um comportamento mais equilibrado dos erros, melhorando assim o desempenho do modelo.

Modelo Regressão Linear Múltipla - log(preço)

Dado isto, criou-se um modelo de regressão linear múltipla novo com o preço em logaritmo. As restantes características do antigo modelo mantiveram-se iguais:

Tabela 1 - Comparação de resultados Regressão Linear Múltipla

Conjunto Teste	Modelo de Regressão Linear Múltipla	Modelo de Regressão Linear Múltipla (Preço Logarítmico)
RMSE	51.16 €	52.05 €
MAE	38.84 €	37.54 €
R2	0.2718	0.2464
Breusch Pagan	750.03	765.97
White	773	783
Breusch-Godfrey	97.192	170.09

Em comparação com o modelo de regressão linear múltipla inicial, o modelo com o preço em logaritmo apresenta um desempenho semelhante na previsão, com um ligeiro aumento do RMSE e uma melhoria do MAE.

Apesar de o R^2 ser mais baixo, o modelo logarítmico garante que todas as previsões são positivas, ou seja, corrige uma limitação importante do modelo linear inicial. Assim, mesmo sem existirem melhorias claras em todas as métricas, este modelo revela-se mais adequado do ponto de vista económico e mais fácil de interpretar.

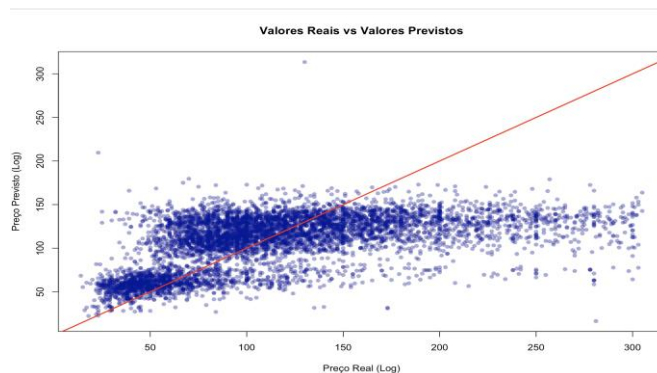


Figura 13 - Valores Reais vs Previstos RLM (log)

O gráfico apresentado compara os valores reais com os valores previstos e mostra que o modelo acompanha a tendência geral dos preços, apesar de existir alguma dispersão em torno da linha de referência. O modelo tem mais precisão nos preços baixos e intermédios, mas tende a subestimar os preços mais elevados, consequentemente as previsões estão num intervalo mais reduzido.

Estes resultados estão de acordo com as métricas obtidas anteriormente, ou seja, o modelo apresenta um desempenho razoável, no entanto tem algumas limitações na previsão de valores mais extremos.

Modelo de Árvore de Decisão (CART)

Foi utilizado um modelo de Árvore de Decisão para regressão, baseado no algoritmo CART, implementado através da função `rpart` com o método *anova*.

Com o objetivo de reduzir o risco de *overfitting* e avaliar a capacidade de generalização do modelo, o processo de treino recorreu a validação cruzada com 10 folds. Neste procedimento, o conjunto de dados de treino é dividido em dez partes aproximadamente iguais, sendo o modelo treinado iterativamente em nove partes e validado na parte remanescente. Este processo permite que todas as observações sejam utilizadas tanto para treino como para validação, assegurando uma avaliação mais estável do desempenho do modelo e apoiando a seleção de uma estrutura de complexidade adequada.

Visualização gráfica do modelo

A Figura 14 apresenta a representação gráfica da árvore de decisão final obtida.

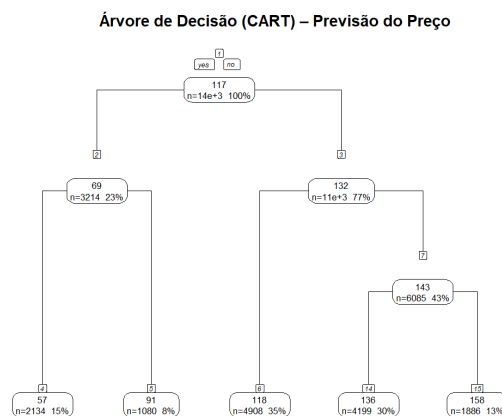


Figura 14 - Árvore de decisão (CART)

Em cada nó da árvore são apresentadas as seguintes informações: o valor médio previsto do preço, o número de observações (n) e a percentagem de observações relativamente ao total. O Nó 1 corresponde ao nó raiz da árvore e apresenta o preço médio global dos anúncios no conjunto de treino. A partir deste nó, a árvore realiza a primeira e mais relevante divisão com base na variável *room_type*, separando os anúncios em dois grandes grupos (Nós 2 e 3).

O Nó 2 (ramo esquerdo) agrega anúncios do tipo *Private room* e *Shared room*, caracterizados por preços médios mais baixos. O Nó 3 (ramo direito) inclui anúncios do tipo *Entire home/apt* e *Hotel room*, associados a preços médios significativamente mais elevados. Esta divisão evidencia que o tipo de alojamento é o fator com maior poder discriminativo no modelo.

No ramo correspondente aos quartos (Nó 2), a árvore introduz uma nova divisão baseada na variável *neighbourhood*, originando dois nós terminais. O Nó 4 agrega anúncios localizados maioritariamente em freguesias de municípios periféricos ou residenciais, como Alenquer, Arruda dos Vinhos, Cadaval, Sobral de Monte Agraço ou Vila Franca de Xira, caracterizados por menor pressão turística e preços médios mais reduzidos. O Nó 5 inclui anúncios localizados em zonas mais centrais ou com maior atratividade turística, como Lisboa (Santa Maria Maior, Misericórdia), Cascais, Oeiras ou Sintra, onde os preços médios dos quartos são substancialmente mais elevados. Assim, para anúncios do tipo quarto, a localização ao nível do município surge como o principal fator explicativo da variação do preço.

No ramo correspondente a alojamentos completos ou quartos de hotel (Nó 3), a árvore volta a recorrer à variável *neighbourhood*, distinguindo dois grandes grupos de freguesias. O Nó 6 inclui freguesias com preços intermédios, como Loures, Odivelas, Amadora ou zonas menos centrais de Sintra. O Nó 7 inclui freguesias com maior centralidade ou procura turística, nos municípios de Lisboa, Cascais, Estoril e áreas mais valorizadas de Sintra.

Neste último subconjunto (Nó 7) verifica-se ainda heterogeneidade adicional, levando à introdução de uma nova divisão. A árvore introduz uma divisão adicional com base na variável *calculated_host_listings_count*, distinguindo dois perfis de anfitrião. O Nó 14 corresponde a anfitriões com menos de 12.5 anúncios ativos, associados a preços médios mais baixos, enquanto o Nó 15 corresponde a anfitriões com 12.5 ou mais anúncios ativos, associados aos preços médios mais elevados da árvore. Esta divisão sugere um efeito de escala ou profissionalização, particularmente relevante em municípios com preços elevados e no segmento de alojamentos completos.

Avaliação de Performance do Modelo

A avaliação do desempenho do modelo combinou a análise gráfica dos valores previstos e dos resíduos com a utilização de métricas quantitativas de erro no conjunto de teste. Nas figuras seguintes, apresentam-se os respetivos resultados.

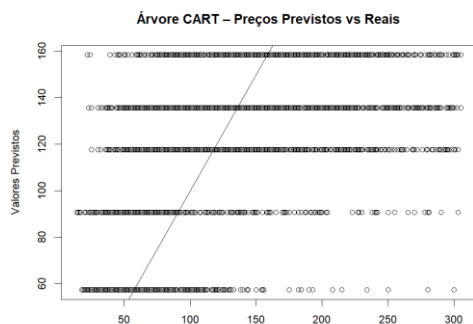


Figura 16 - Preços previstos vs Preços Reais (Árvore de Decisão)

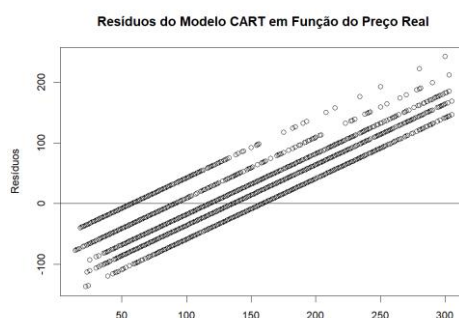


Figura 15 - Resíduos Árvore de Decisão

O gráfico de valores previstos face aos valores reais mostra que o modelo capta a tendência global dos dados, mas produz previsões discretas, correspondentes às médias dos nós terminais da árvore. Esta característica reflete a estrutura do modelo, que segmenta os dados em um número limitado de regiões homogêneas.

A análise dos resíduos evidencia padrões sistemáticos, nomeadamente um aumento do erro para valores de preço mais elevados, indicando limitações na capacidade do modelo em capturar variações mais finas do preço dentro de cada segmento.

As métricas de erro obtidas foram $MAE = 39.14295$ e $RMSE = 51.3437$. O valor do MAE indica que, em média, as previsões do modelo afastam-se cerca de 39,14€ dos valores reais. Por sua vez, o valor mais elevado do RMSE, cerca de 51,34€, evidencia a presença de previsões com erros de maior magnitude associados a anúncios com preços mais elevados.

Análise da complexidade do modelo e modelo podado

A complexidade da árvore de decisão foi analisada através da tabela e do gráfico que relacionam o erro relativo de validação cruzada com o parâmetro de complexidade (cp).

	CP	nsplit	rel error	xerror	xstd
1	0.189336	0	1.00000	1.00024	0.012788
2	0.032894	1	0.81066	0.81088	0.011649
3	0.015255	2	0.77777	0.78333	0.011541
4	0.013005	3	0.76251	0.77250	0.011255
5	0.010000	4	0.74951	0.75739	0.011214

Figura 17 - Tabela de complexidade

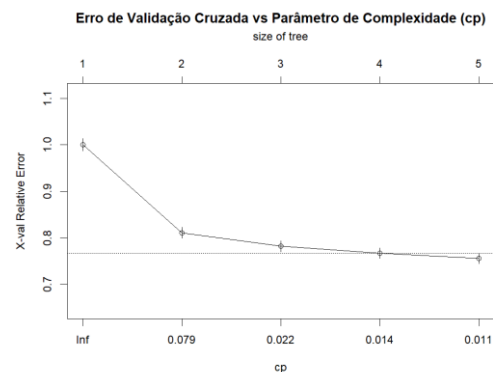


Figura 18 - Erro de validação cruzada vs complexidade

Observa-se que o erro diminui à medida que a complexidade do modelo aumenta, embora os ganhos se tornem progressivamente marginais a partir de determinado ponto. O menor valor do erro de validação cruzada foi identificado para $cp \approx 0.01$, o que motivou a seleção deste valor para esta etapa.

Com base neste valor, foi construída uma versão podada da árvore de decisão. A comparação entre a árvore original e a árvore podada revela que a estrutura do modelo não apresentou alterações estruturais, indicando que a árvore inicial já se encontrava num nível adequado de complexidade. Consequentemente, as métricas de desempenho no conjunto de teste também permaneceram inalteradas, confirmando que a aplicação da poda não produziu melhorias adicionais no desempenho preditivo.

Modelo Bagging

Foi desenvolvido um modelo baseado em Bagging (Bootstrap Aggregating), recorrendo ao método *treebag* do pacote *caret*. O treino do modelo foi realizado com validação cruzada a 10 folds e foi construído a partir da agregação de 100 árvores de decisão ($nbagg = 100$). Esta abordagem tem como objetivo reduzir a variância associada a uma única árvore, através da agregação das previsões de múltiplos modelos treinados sobre diferentes amostras do conjunto de dados.

Visualização gráfica do modelo

Embora o modelo Bagging resulte da combinação de várias árvores de decisão, é possível visualizar as árvores individuais com o objetivo de compreender o tipo de regras aprendidas pelo modelo. A árvore apresentada deve, contudo, ser interpretada apenas de forma ilustrativa, uma vez que a previsão final do modelo resulta da agregação das previsões de todas as árvores.

Optamos por visualizar a árvore número 10, apresentada na figura seguinte.

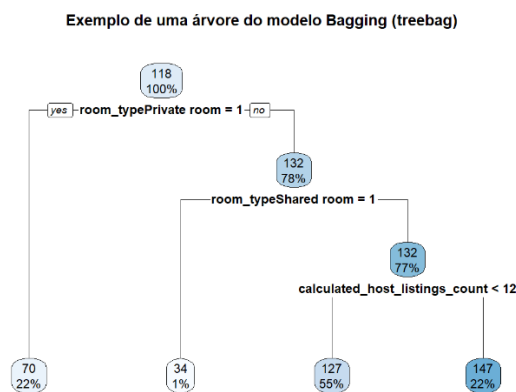


Figura 19 - Exemplo de árvore de decisão (Bagging)

Na árvore apresentada, observa-se que as primeiras divisões recorrem novamente à variável *room_type*, separando anúncios do tipo *Private room* dos restantes. Em níveis subsequentes surgem divisões associadas a *Shared room* e à variável *calculated_host_listings_count*, confirmando a relevância destas variáveis na explicação do preço. Esta estrutura é consistente com as divisões identificadas no modelo CART, sugerindo estabilidade nos principais fatores dos modelos.

Avaliação de Performance do Modelo

O desempenho do modelo Bagging foi avaliado através da análise gráfica dos valores previstos e de métricas quantitativas de erro.

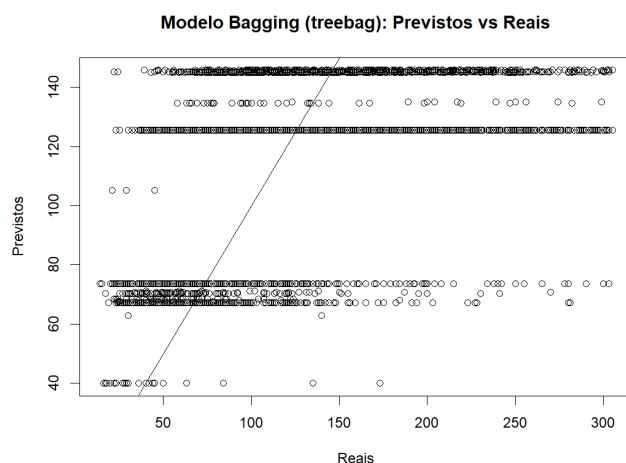


Figura 20 - Preços previstos vs reais (bagging)

O gráfico de valores previstos face aos valores reais mostra que o modelo acompanha a tendência global dos dados, mas apresenta previsões concentradas em patamares horizontais, comportamento típico de métodos baseados em árvores. Observa-se ainda uma maior dispersão

relativamente à linha de referência para preços mais elevados, indicando dificuldades na modelação dos extremos da distribuição.

As métricas de erro obtidas no conjunto de teste foram $MAE = 40.5669$ e $RMSE = 52.74868$. Em comparação com o modelo CART, observa-se um aumento em ambas as métricas, indicando que o modelo Bagging apresenta, em média, previsões menos próximas dos valores reais e uma maior incidência de erros de maior magnitude.

Importância das variáveis e comparação com o modelo CART

A Figura 21 apresenta o gráfico de importância das variáveis para o modelo Bagging, considerando as 15 variáveis mais relevantes para a previsão do preço. Este gráfico permite identificar os atributos que mais contribuem para a redução do erro ao longo das várias árvores que compõem o ensemble.

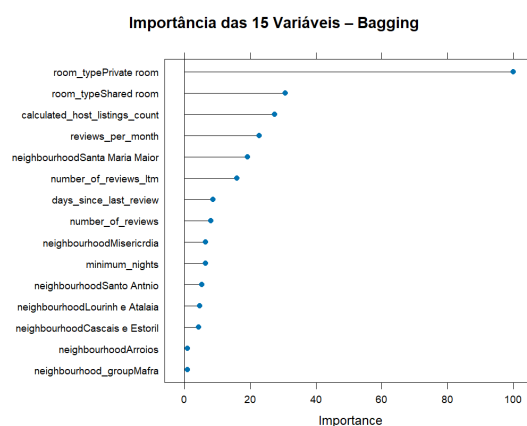


Figura 21 - Variáveis mais importantes (bagging)

Observa-se que variáveis relacionadas com o tipo de alojamento, em particular *room_type*, surgem entre as mais importantes, confirmando o papel central desta variável na diferenciação dos preços. De forma semelhante, a variável *calculated_host_listings_count* apresenta elevada relevância, evidenciando novamente o impacto da escala de atividade do anfitrião na formação do preço. Adicionalmente, variáveis associadas à localização e à dinâmica de procura, como indicadores de disponibilidade ao longo do ano e número de avaliações, também surgem entre as mais relevantes, sugerindo que o modelo combina informação estrutural com sinais de procura.

Quando comparado com o modelo CART, verifica-se que os principais fatores explicativos do preço permanecem consistentes entre metodologias. No entanto, esta consistência não se traduziu numa melhoria do desempenho preditivo, uma vez que o modelo Bagging apresentou valores de MAE e RMSE ligeiramente superiores. Assim, embora esta técnica proporcione maior robustez e estabilidade das previsões, esta não foi suficiente para superar o desempenho do modelo de Árvore de Decisão simples neste conjunto de dados.

Modelo Random Forest

O modelo de Random Forest (Florestas Aleatórias) é um modelo que utiliza várias árvores de decisão para melhorar previsões. Cada árvore analisa uma parte aleatória dos dados e os seus resultados são combinados através de uma votação majoritária (bagging) para a classificação ou através da média para a regressão.

Principais parâmetros do modelo

Os principais parâmetros do modelo Random Forest utilizados foram:

- **mtry**: número de variáveis consideradas aleatoriamente em cada divisão da árvore.
- **splitrule**: critério utilizado para escolher a melhor divisão.
- **min.node.size**: número mínimo de observações permitidas em cada nó terminal
- **num.trees**: número total de árvores na floresta
- **importance**: método de cálculo da importância das variáveis.

Modelo Random Forest (sem Ajuste Manual)

Descrição do método

Numa primeira fase foi usado o método ranger da biblioteca caret, o modelo correu sem definições explícitas apenas sendo usado o tuneLength, que testa diferentes combinações de parâmetros. Nesta fase inicial usamos duas abordagens diferentes para a validação cruzada, uma validação com 10 folds e outra com 5 folds.

Avaliação de Performance do Modelo

Tabela 2 - Comparação de resultados folds

Conjunto Teste			RMSE	Rsquared	MAE
folds	Splirule	mtry			
10	variance	2	56.86152	0.2555144	44.98411
10	extratrees	2	56.96352	0.2339881	45.05862
10	variance	38	48.36436	0.3631244	36.10547
10	extratrees	38	49.22335	0.3401756	36.87317
10	variance	74	48.19719	0.3656987	35.64648
10	extratrees	74	48.13621	0.3674078	35.52759
10	variance	110	48.27199	0.3639048	35.65047
10	extratrees	110	48.08835	0.3689186	35.38143
10	variance	147	48.40284	0.3607590	35.73393
10	extratrees	147	48.05020	0.3701420	35.31436
5	variance	2	56.88437	0.2576075	45.00272
5	extratrees	2	56.93379	0.2318812	45.05580
5	variance	38	48.50558	0.3594346	36.26115
5	extratrees	38	49.38539	0.3356684	37.05553
5	variance	74	48.39725	0.3606532	35.91625
5	extratrees	74	48.38816	0.3609163	35.83223

5	variance	110	48.47440	0.3588925	35.91637
5	extratrees	110	48.35216	0.3622424	35.70210
5	variance	147	48.60208	0.3559409	35.98283
5	extratrees	147	48.34992	0.3625993	35.67757

Resultados com validação cruzada a 10 folds

A partir da tabela podemos observar com uma validação cruzada de 10 folds, o modelo testou várias combinações dos parâmetros `mtry` e `splitrule`, mantendo `min.node.size = 5`.

O melhor modelo apresentou os seguintes parâmetros: `mtry = 147` e `splitrule = extratrees`. Este modelo obteve um RMSE de 48.05020, indicando que o modelo apresenta, em média, um erro de aproximadamente 48€, sendo esta métrica mais sensível a erros grandes. O R^2 de 0.3701420 mostra que o modelo consegue explicar cerca de 37% da variabilidade da variável resposta.

Já o MAE, embora não seja o melhor entre todos os modelos, o valor de 35.31436 indica que o erro médio absoluto do modelo é de cerca de 35€, refletindo o erro típico das previsões.

Resultados com validação cruzada a 5 folds

Tal como o resultado com a validação cruzada de 10 folds também a validação cruzada a 5 folds, de forma semelhante, indicou o mesmo conjunto de parâmetros como sendo o melhor.

O melhor modelo apresentou os seguintes parâmetros: `mtry = 147` e `splitrule = extratrees`. Este modelo obteve um RMSE de 48.34992, indicando que o modelo apresenta, em média, um erro de aproximadamente 48 unidades, sendo esta métrica mais sensível a erros grandes. O R^2 de 0.3625993 mostra que o modelo consegue explicar cerca de 36% da variabilidade da variável resposta.

Já o MAE, embora não seja o melhor entre todos os modelos, o valor de 35.67757 indica que o erro médio absoluto do modelo é de cerca de 36€, refletindo o erro típico das previsões.

Comparando estes valores com a validação cruzada de 10 folds, podemos observar que os resultados são muito semelhantes, isto indica que o modelo não é significativamente influenciado pelo número de folds utilizados.

Podemos concluir que a utilização de 10 folds permite uma avaliação ligeiramente mais rigorosa do desempenho do modelo, enquanto 5 folds oferece resultados praticamente equivalentes com maior eficiência computacional.

É de ressaltar que devido à complexidade do modelo não foi possível fazer mais testes, pois o tempo necessário para cada teste varia de 1h-2h.

Modelo Random Forest com Ajuste dos Hiperparâmetros

Depois de verificarmos que os modelos anteriores tinham um *runtime* elevado optamos por

fazer um modelo Random Forest onde fizemos um ajuste manual dos hiperparâmetros.

Foram testadas as seguintes combinações:

- mtry: 3, 4, 5, 6
- min.node.size: 5, 10 e 20
- splitrule: variance e extratrees
- num.trees: 500 árvores e 1000 árvores

Resultados do Modelo Ajustado

Tabela 3 - Comparação 2 melhores resultados modelo ajustado

Conjunto Teste			RMSE	Rsquared	MAE	num.trees
folds	splitrule	mtry				
6	variance	10	52,44315	0,3021	40,80692	500
6	variance	5	53,31148	0,2563	41,51892	500

A partir da comparação direta entre os melhores modelos para cada *splitrule*, conclui-se que o modelo com o critério variance apresenta o melhor desempenho global. Este critério resulta no menor RMSE, no menor MAE e na maior capacidade explicativa, medida pelo coeficiente de determinação (R^2).

Por outro lado, o modelo que utiliza o *splitrule* extratrees revela um desempenho inferior, apresentando valores piores em todas as métricas avaliadas, o que indica menor precisão e menor capacidade de explicação da variabilidade dos dados.

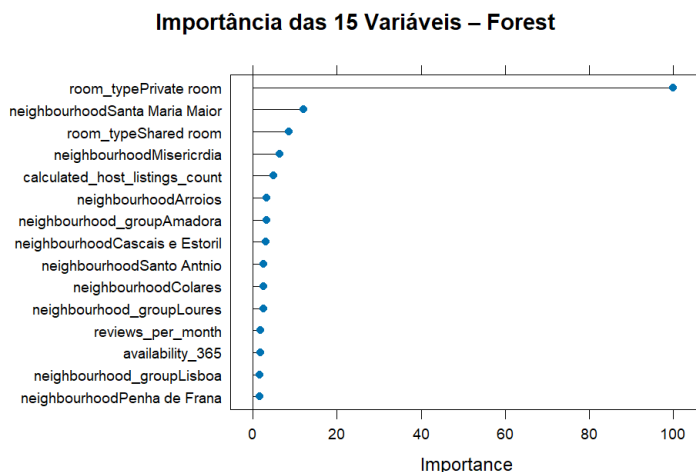


Figura 22 - Importância das variáveis Random Forest

Modelo Gradient Boosting Machine (GBM)

De seguida procedemos com a criação de um modelo GBM (Gradient Boosting Machine), que pode ser entendido como uma evolução das técnicas do modelo Boosting, nas quais os modelos base são construídos de forma sequencial. Em cada iteração, o novo modelo é ajustado tendo em vista os erros cometidos pelos modelos anteriores, procurando corrigi-los e melhorar

progressivamente a capacidade preditiva do conjunto.

Avaliação da performance do modelo

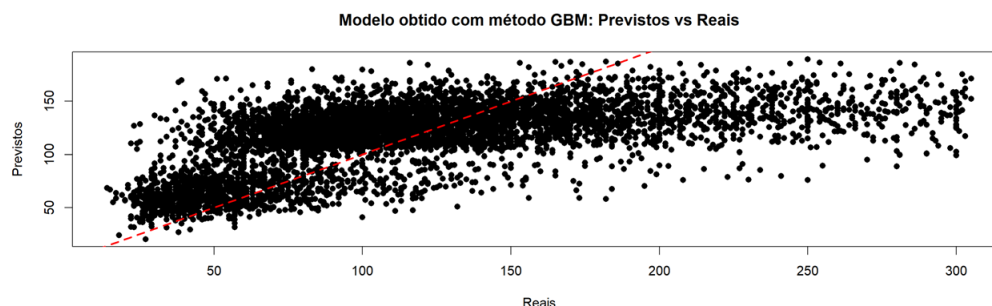


Figura 23 - Preços previstos vs Reais (GBM)

Através do gráfico obtido podemos concluir que o modelo consegue prever com maior assertividade os preços dos quartos mais baratos do que os quartos mais caros.

As métricas de erro obtidas foram $MAE = 36.6215$ e $RMSE = 48.28227$. O valor do MAE indica que, em média, as previsões do modelo afastam-se cerca de 36,62€ dos valores reais. Por sua vez, o valor mais elevado do RMSE, cerca de 48,28€.

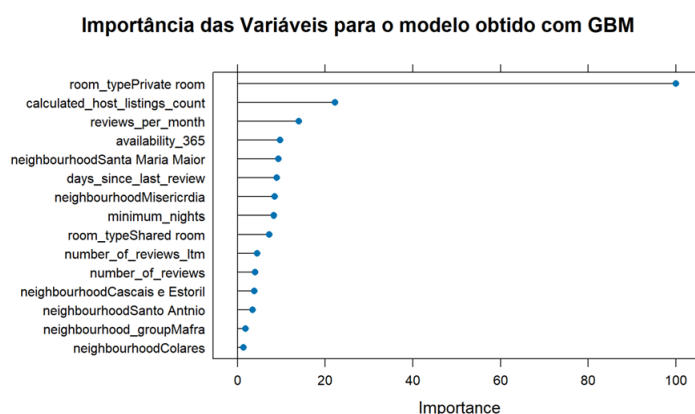


Figura 24 - Top 15 Importância das Variáveis (GBM)

Modelo GBM com ajuste de parâmetros

De forma a otimizar o desempenho do modelo, testamos dois outros modelos com parâmetros ajustados para equilibrar trade-off entre o enviesamento e a variância, sendo que os resultados foram os seguintes:

Tabela 4 - Comparação de Modelo GBM inicial vs ajuste de parâmetros

Modelo	Parâmetros	RMSE	MAE	Feedback
Teste 1 – Modelo Inicial	-	48.28€	36.62 €	Melhor RMSE; Bom equilíbrio entre erro médio e erro absoluto.
Teste 2 – Ajustamento 1	n.trees: 100, 150, 180 e 200; n.minobsinnode: 5 e 10; Interaction.depth: 2; shrinkage: 0.1;	49.92 €	37.79 €	Desempenho inferior em ambas as métricas; Não recomendado
Teste 3 – Ajustamento 2	n.trees: 200 e 300; n.minobsinnode: 5;	48.50 €	36.48 €	RMSE semelhante ao melhor modelo; Melhor MAE

	Interaction.depth: 2 e 3; shrinkage: 0.05 e 0.1;			
--	---	--	--	--

Podemos concluir que os resultados do “Ajustamento 2” foram ligeiramente melhores que os resultados do “Ajustamento 1”. No entanto, ao compararmos com os resultados obtidos com o teste inicial, podemos interpretar que os resultados obtidos do MAE foram melhores, no entanto, relativamente ao RMSE foram ligeiramente piores.

Ao avaliar o desempenho do modelo de GBM, podemos concluir que, apesar da otimização, ele apresenta limitações na precisão das previsões. Dado que o preço médio dos quartos é de 117€, o RMSE apresenta um desvio de 41.3% do preço médio. Este nível de erro é elevado e sugere que, ou a variância nos preços de Lisboa é muito grande, ou que o conjunto de variáveis preditivas utilizado pode não ser suficiente para capturar todos os fatores que influenciam os preços dos quartos.

Conclusão

Com a realização destes diferentes modelos de previsão, conseguimos colocar em prática os modelos de previsão aplicados a problemas de regressão explorados em contexto de aula.

Apesar da aplicação das aprendizagens de aula, ao contrário dos casos explorados, a aplicação de modelos mais complexos, como árvores de decisão e as correspondentes técnicas de melhoria de desempenho (bagging, random forest, boosting) acabaram por não se traduzir em ganhos de previsão significativos, face à regressão linear múltipla aplicada inicialmente.

A dificuldade em alcançar reduções de RMSE e MAE com o aumento da complexidade dos modelos poderá ser em parte caracterizada pelas variáveis do dataset original. A falta de correlações fortes (particularmente nas variáveis numéricas) com a variável preço tornou assim as previsões mais fracas, especialmente por não termos acesso a todo um conjunto de informação relevante para a previsão dos preços (como nº de quartos, casas de banho – privativa/partilhada, nº de hóspedes máximo, dimensão do espaço, segmento – luxo, económico...).

Mesmo com estas dificuldades, conseguimos observar um conjunto de variáveis relevantes para a previsão de preço, destacando-se principalmente o *room_type* (particularmente *private* e *shared room*, com preços significativamente mais baixos que *hotel room* e *entire room/apt*), e *neighbourhood* (destacando-se as freguesias onde o alojamento local é mais comum – baixa lisboeta). No âmbito de estudos posteriores poderia ser relevante analisar ainda o nome dos anúncios, pois poderia ser uma característica relevante para a previsão de preços, por via de *text mining* ou outras ferramentas.

Tendo em vista todos estes fatores, o modelo que alcançou a melhor performance foi o modelo de florestas aleatórias, com uma validação cruzada de 10 folds, *mtry* = 147 e *splitrule* = *extratrees*. Este modelo obteve um RMSE de 48.05020, R^2 de 0.3701420 e um MAE de 35.31436, apesar de ter a clara característica de garantir previsões discretas para uma variável contínua, como é de esperar dos modelos de árvores de decisão.

Este trabalho demonstrou as dificuldades inerentes à previsão de preços no mercado do alojamento local, onde a quantidade de dados é elevada, mas nem todos os dados apresentam influência sobre os preços, resultando em modelos mais complexos que nem sempre garantem melhorias substanciais.