

APRENDIZADO DE MÁQUINA

DANIEL DOS SANTOS E LYNCOLN SOUSA

15/09/2020

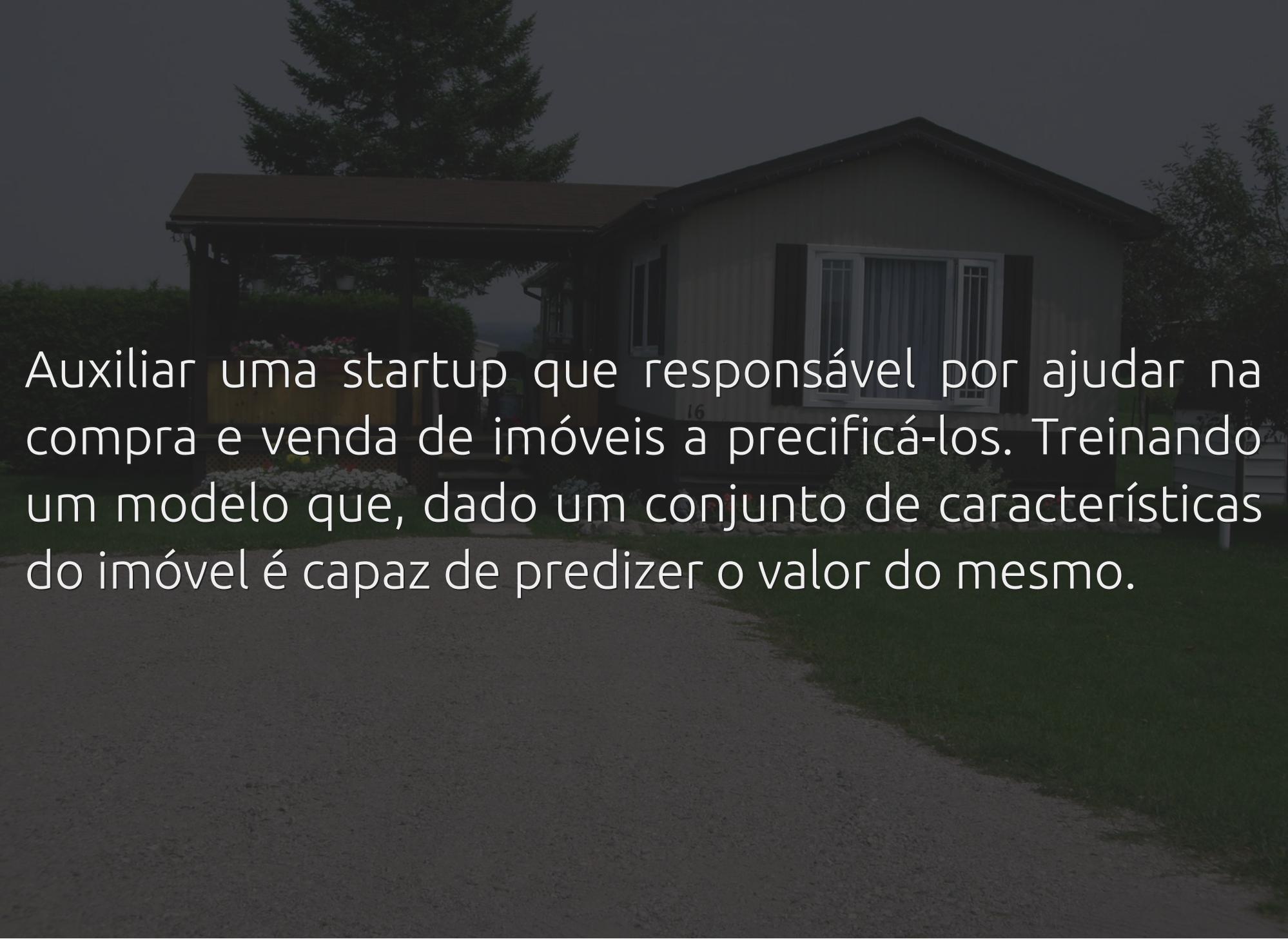
CONTEÚDO

- O problema
- Metodologia
- Resultados
- Conclusão

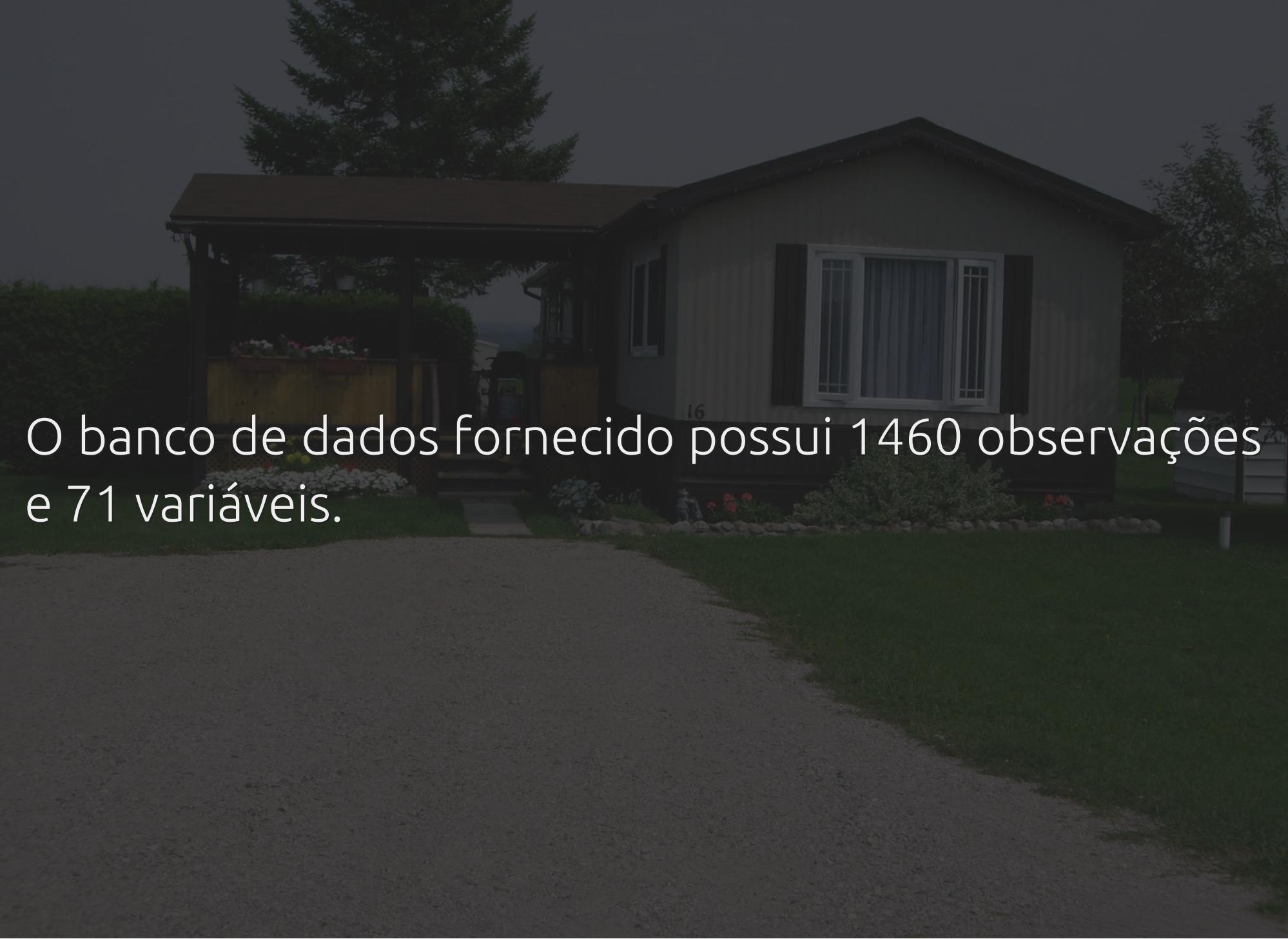


O PROBLEMA

16



Auxiliar uma startup que responsável por ajudar na compra e venda de imóveis a precificá-los. Treinando um modelo que, dado um conjunto de características do imóvel é capaz de predizer o valor do mesmo.



O banco de dados fornecido possui 1460 observações
e 71 variáveis.

METODOLOGIA

Uma breve análise exploratória foi realizada antes de qualquer outro processo.

PRÉ-PROCESSAMENTO

Para lidar com os dados faltantes da variável **LotFrontage**, tamanho da rua conectada à propriedade, foi transformada em uma categoria que separa os tamanhos em 4:

São estes, de **0 a 60** pés, **de 60 a 110** pés, **maior que 110** pés e **NA** que indica que o imóvel não possui uma rua conectada.

O mesmo foi feito com a variável **MasVnrArea** porém utilizando 3 categorias:

Área de alvenaria de **0 a 100** pés quadrados, **maior que 100** pés quadrados e **NA** que representa a falta de área de alvenaria.

A variável que contém o ano em que a garagem foi construída, nomeada de **GarageYrBlt**, sofreu uma transformação que calcula a idade da garagem considerando o ano atual. Para os dados faltantes, estes foram considerados **0 (zero)**.

Após essas modificações foram retiradas variáveis com variância próximas ou iguais a zero. Além disso, também verificou-se a existência de variáveis correlacionadas.

Por fim, devido ao grande número de variáveis, foi realizado um PCA. Criando 18 componentes que juntas capturam pelo menos 95% da variação.

O modelo **Gradient Boosting** foi escolhido e serão utilizadas diversas combinações de hiperparâmetros, visando os modelos com menor **RMSE**(raiz quadrada do erro quadrático médio). Uma breve descrição dos hiperparâmetros:

O modelo **Gradient Boosting** foi escolhido e serão utilizadas diversas combinações de hiperparâmetros, visando os modelos com menor **RMSE**(raiz quadrada do erro quadrático médio). Uma breve descrição dos hiperparâmetros:

- **interaction.depth**: A profundidade da árvore;
- **n.trees**: Número de árvores;
- **shrinkage**: É a taxa de aprendizado do modelo;
- **n.minobsinnode**: Número de nós.

Foram utilizadas duas formas de reamostragem:

- **Bootstrap**: Com 10 repetições;
- **K-fold repetido**: Com 10 folds e 3 repetições.

A tabela abaixo mostra para quais as combinações estes hiperparâmetros serão treinados:

A tabela abaixo mostra para quais as combinações estes hiperparâmetros serão treinados:

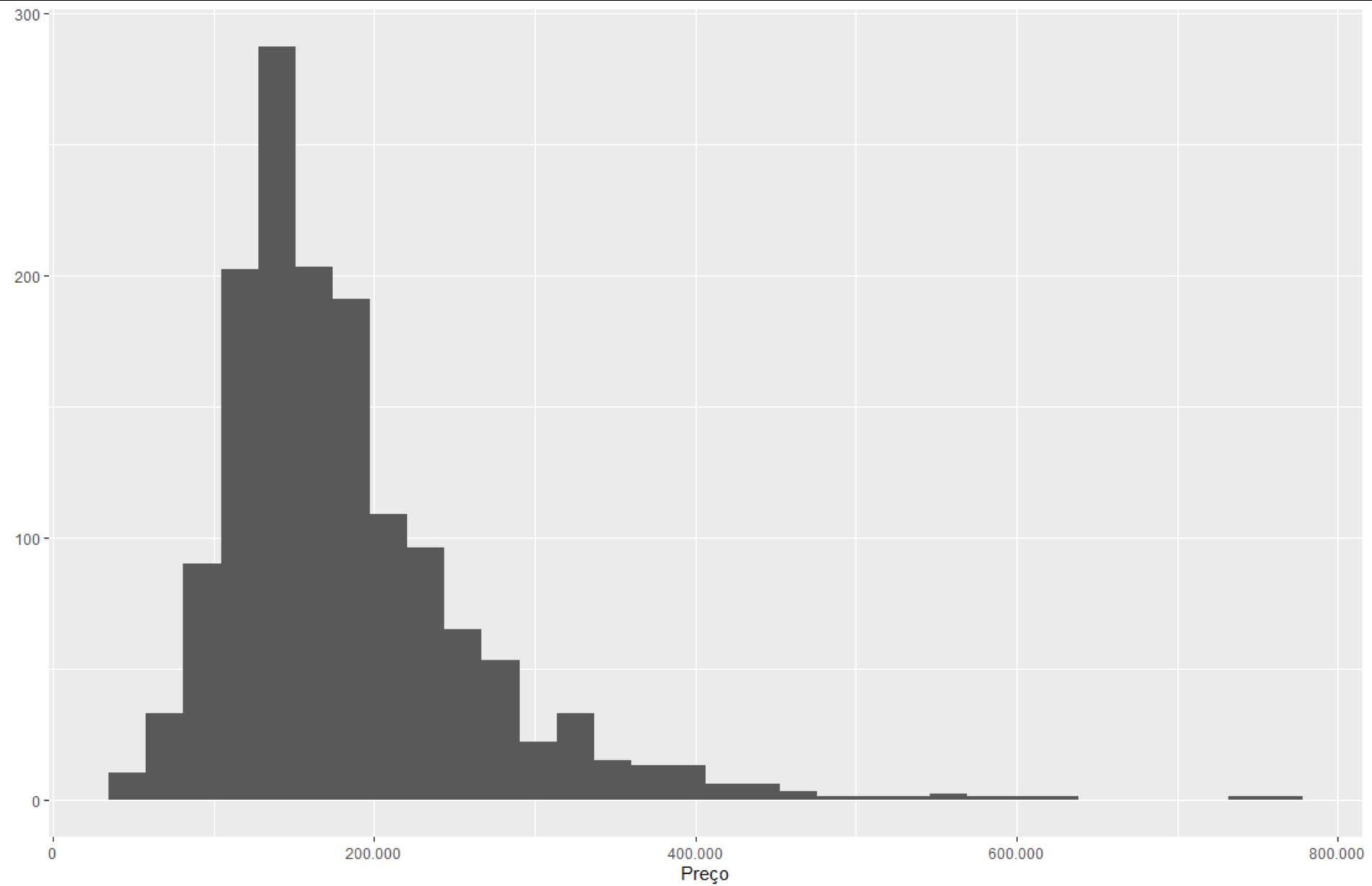
n.minobsinnode	10							
shrinkage	0.01	0.02	0.03					
n.trees	500	1000	1500					
interaction.depth	1	2	3	5	8	13		

Os primeiros modelos a serem treinados com estas combinações de parâmetros possuíam variáveis com baixíssima influência relativa.

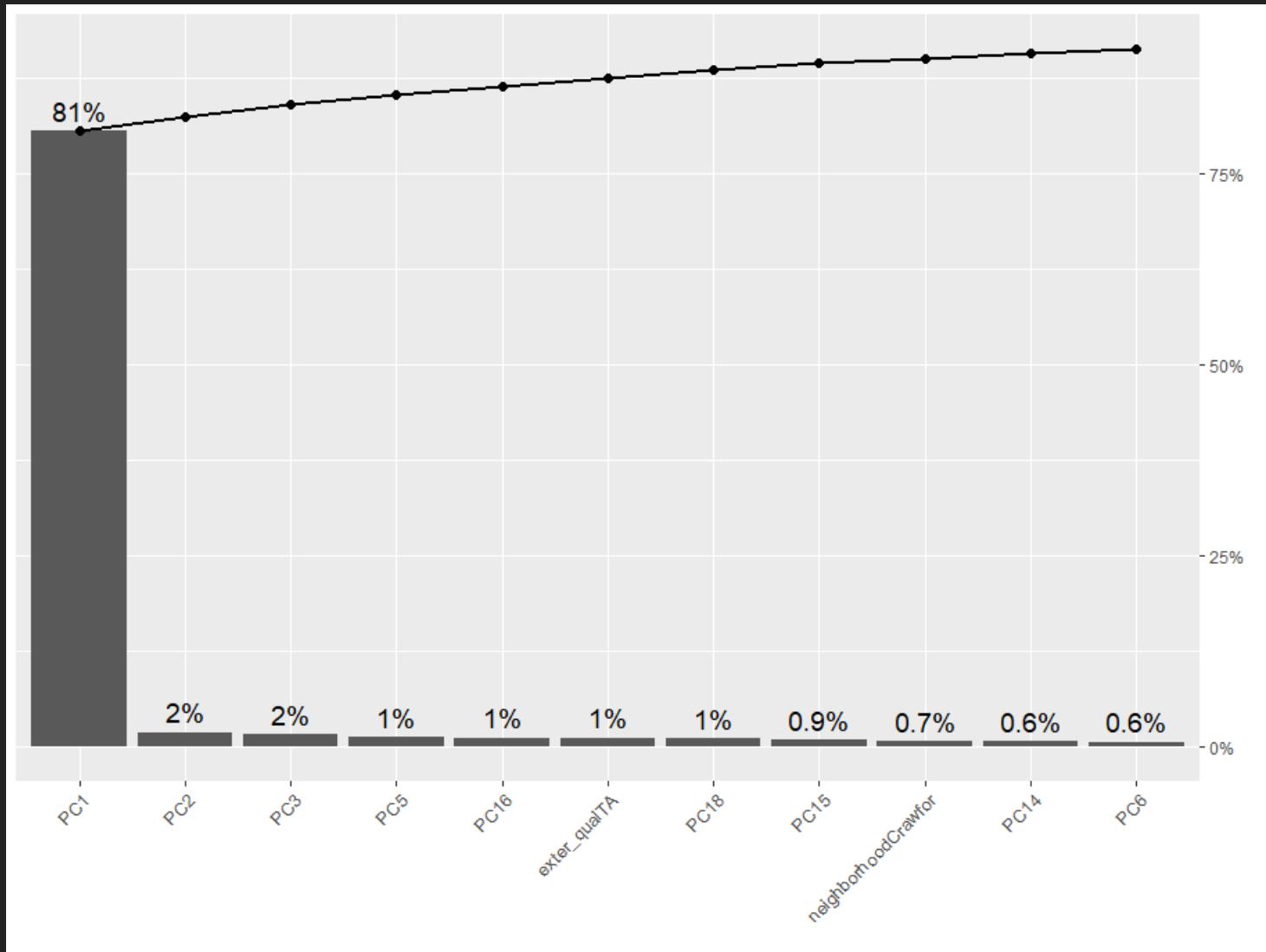
Foram mantidas as variáveis cuja a soma acumulada de suas influências relativas pelo **Gradient Boosting** totalizava 95% e as componentes geradas pelo PCA.

RESULTADOS

Histograma dos preços dos imóveis

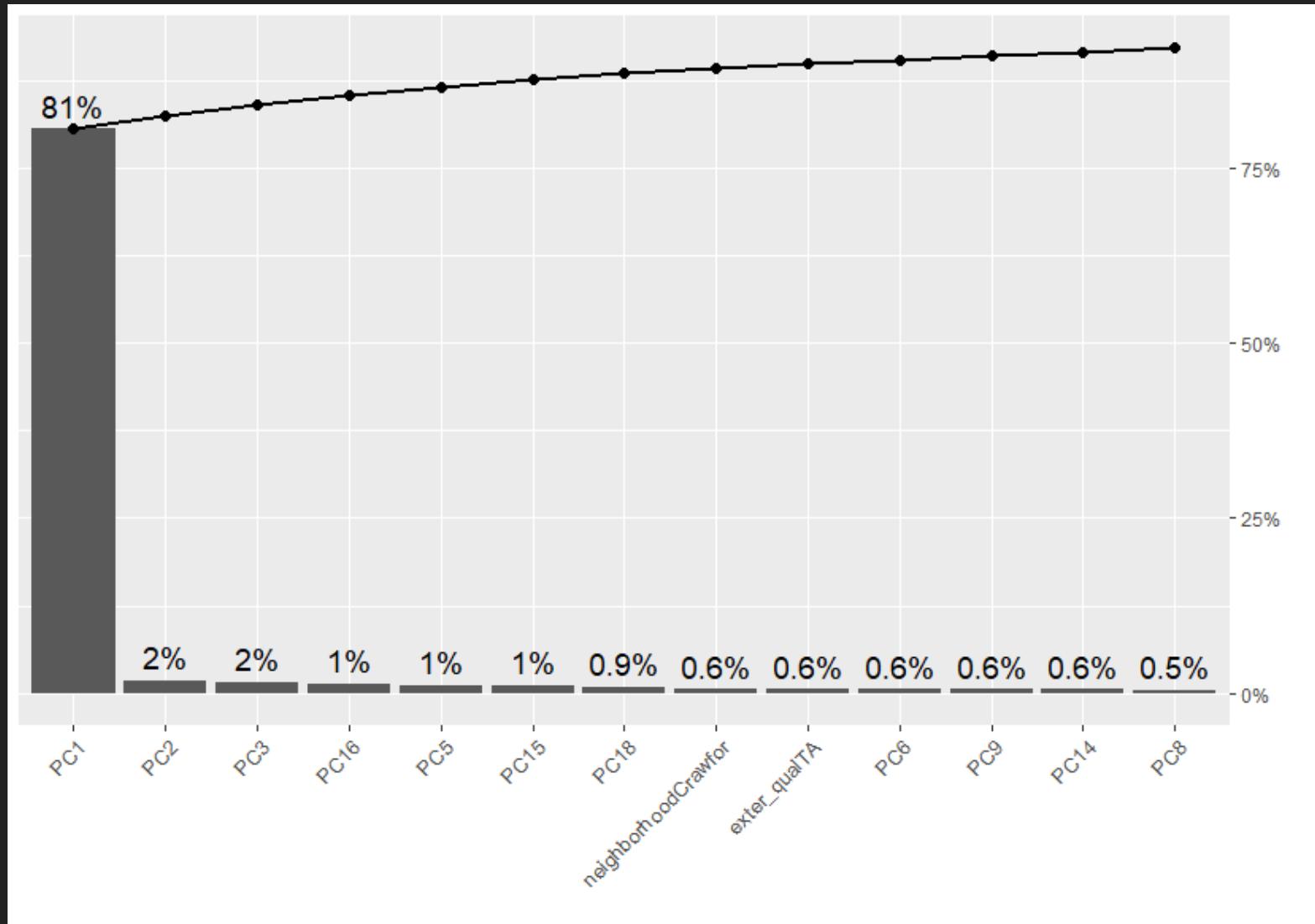


Influência relativa acumulada (Bootstrap)



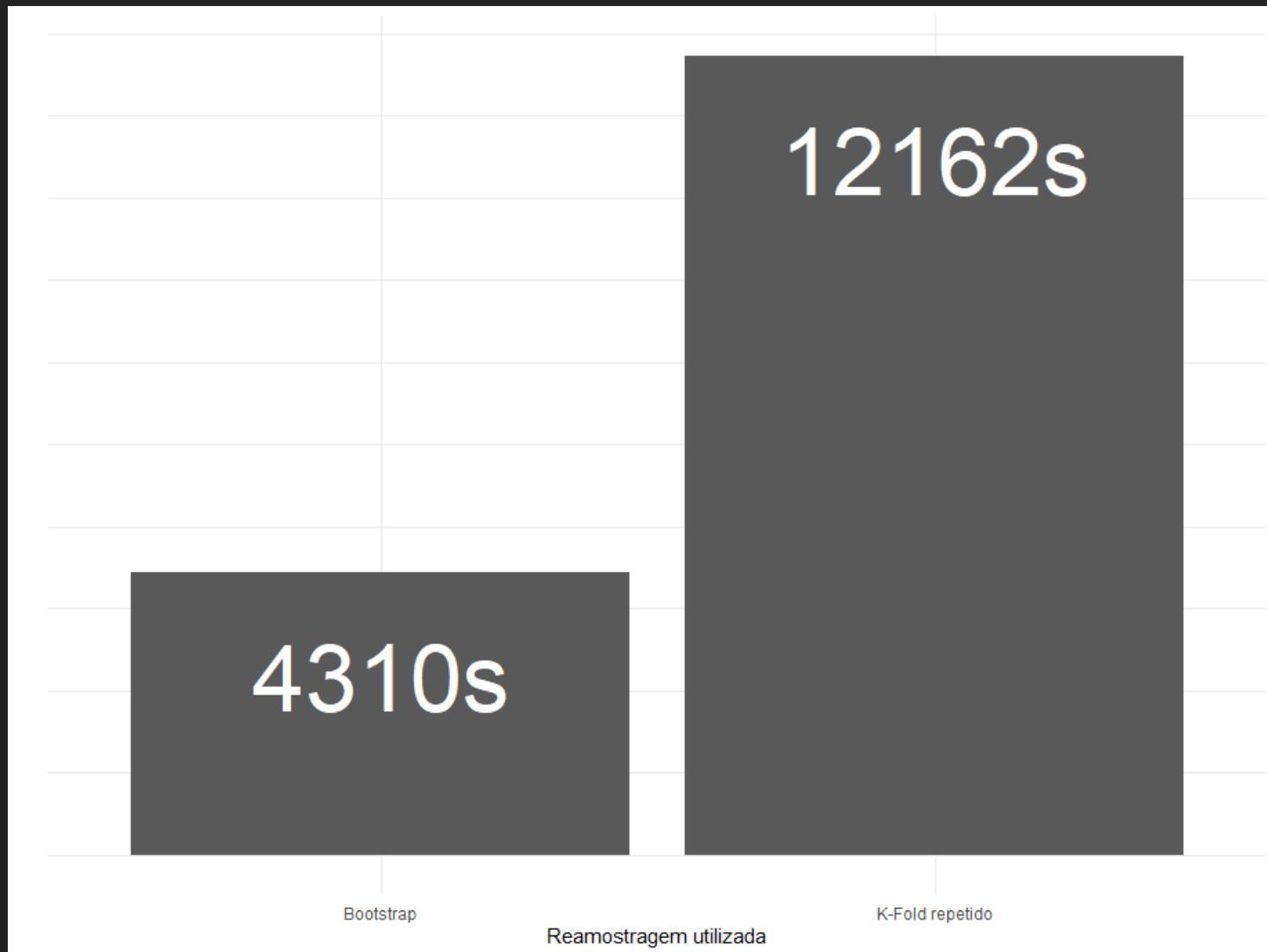
Variáveis com menos de 0.05 de influência relativa estão ocultadas

Influência relativa acumulada (K-Fold repetido)

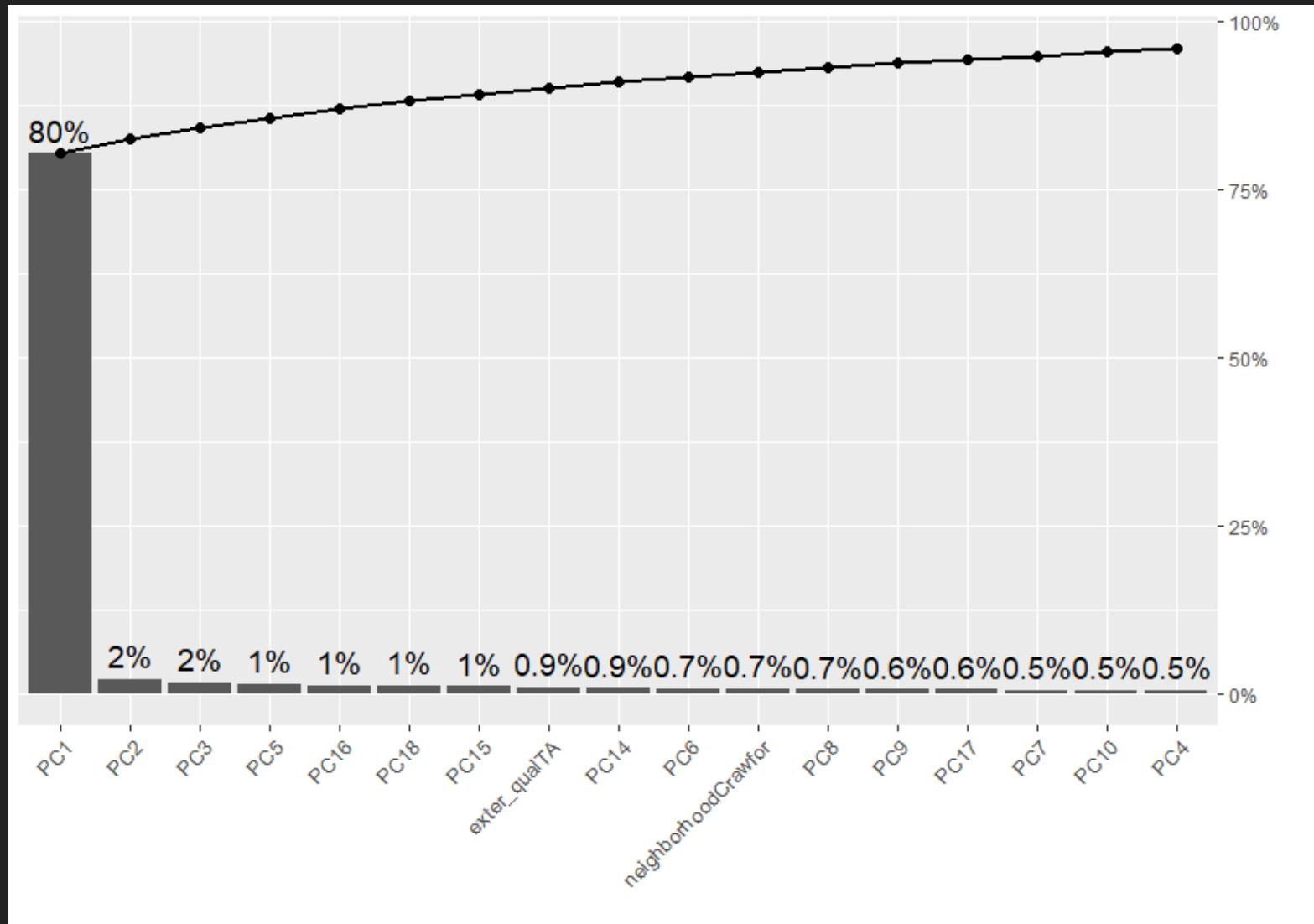


Variáveis com menos de 0.05 de influência relativa estão ocultadas

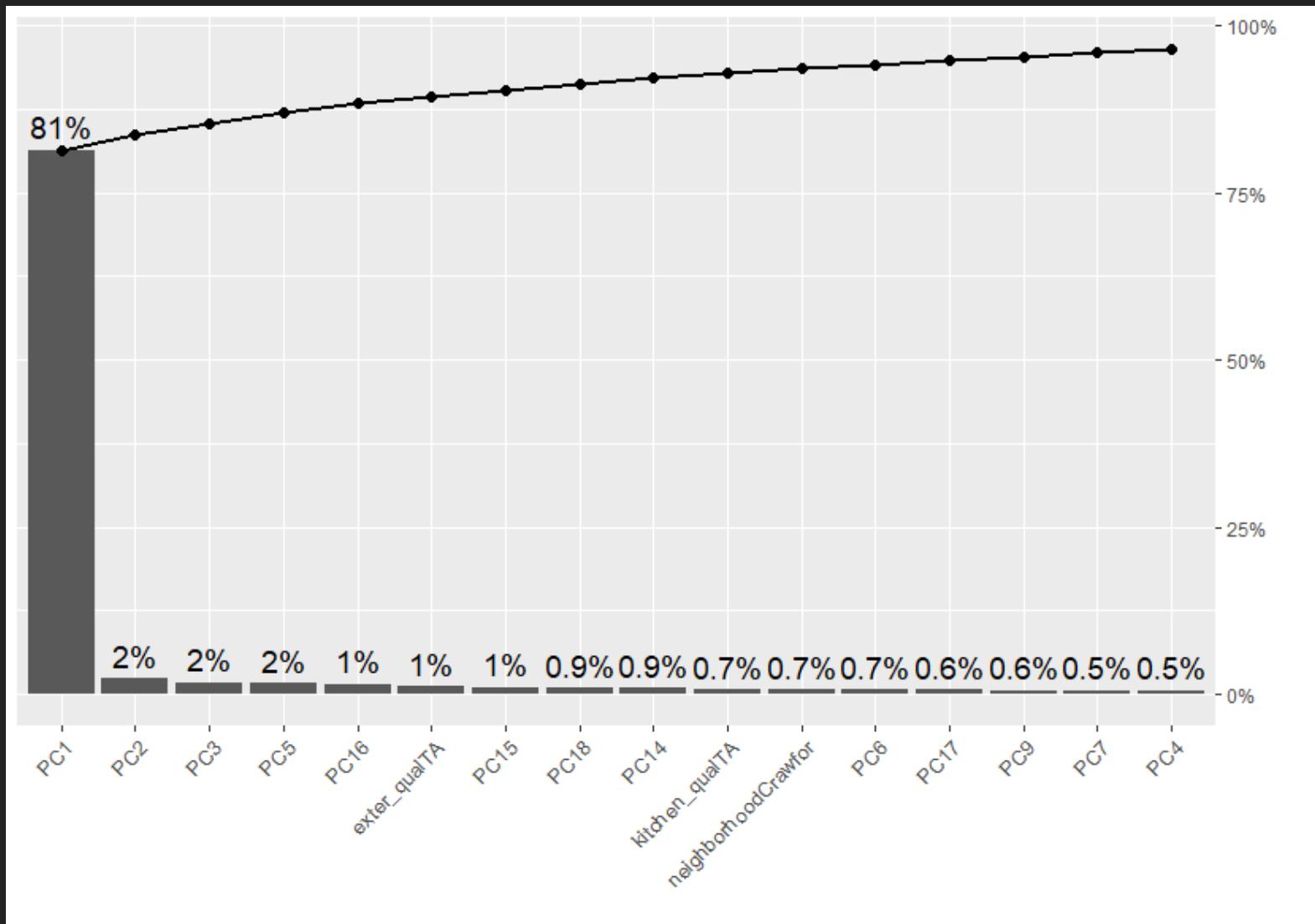
Tempo de treinamento dos modelos



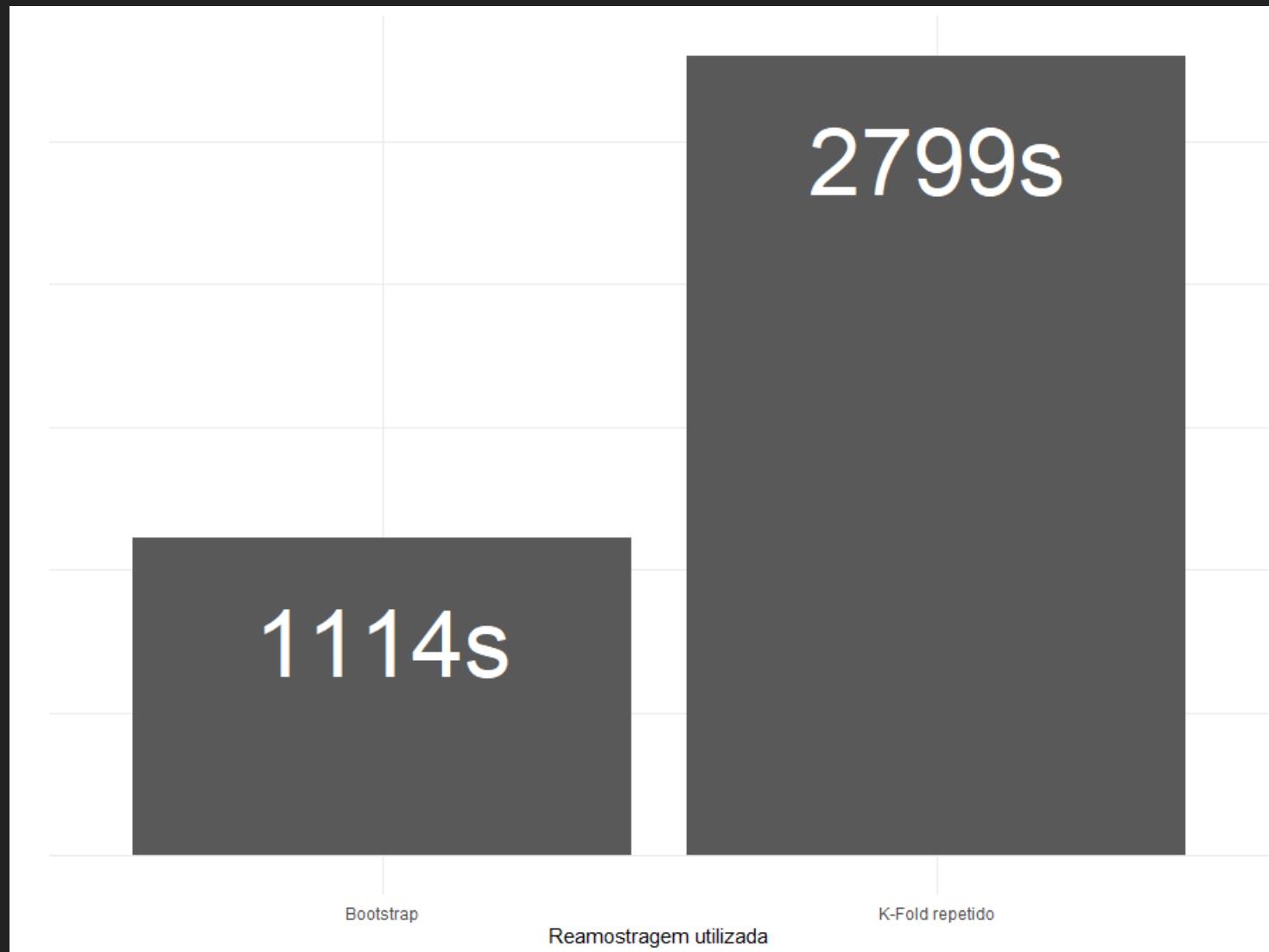
Influência relativa acumulada (Bootstrap - modelo reduzido)



Influência relativa acumulada (K-Fold repetido - modelo reduzido)



Tempo de treinamento dos modelos reduzidos



Antes da redução de variáveis os modelos que se destacaram foram os que apresentaram os seguintes parâmetros:

BOOTSTRAP

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.02	5	10	1500	27836.28	0.8788024	16489.00	3937.117	0.0368602	921.0254
0.03	5	10	1500	27848.94	0.8786265	16581.48	4128.733	0.0392156	956.8659
0.02	8	10	1500	27913.18	0.8778608	16462.09	3964.889	0.0387536	892.2706
0.03	5	10	1000	27947.16	0.8778602	16656.20	4061.951	0.0386882	949.9610
0.02	8	10	1000	28014.23	0.8770834	16522.78	3901.780	0.0381304	895.6395
0.02	5	10	1000	28036.07	0.8771401	16642.99	3850.130	0.0362019	902.3398

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.02	5	10	1500	27836.28	0.8788024	16489.00	3937.117	0.0368602	921.0254
0.03	5	10	1500	27848.94	0.8786265	16581.48	4128.733	0.0392156	956.8659
0.02	8	10	1500	27913.18	0.8778608	16462.09	3964.889	0.0387536	892.2706
0.03	5	10	1000	27947.16	0.8778602	16656.20	4061.951	0.0386882	949.9610
0.02	8	10	1000	28014.23	0.8770834	16522.78	3901.780	0.0381304	895.6395
0.02	5	10	1000	28036.07	0.8771401	16642.99	3850.130	0.0362019	902.3398

K-FOLD REPETIDO

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.02	5	10	1500	27836.28	0.8788024	16489.00	3937.117	0.0368602	921.0254
0.03	5	10	1500	27848.94	0.8786265	16581.48	4128.733	0.0392156	956.8659
0.02	8	10	1500	27913.18	0.8778608	16462.09	3964.889	0.0387536	892.2706
0.03	5	10	1000	27947.16	0.8778602	16656.20	4061.951	0.0386882	949.9610
0.02	8	10	1000	28014.23	0.8770834	16522.78	3901.780	0.0381304	895.6395
0.02	5	10	1000	28036.07	0.8771401	16642.99	3850.130	0.0362019	902.3398

K-FOLD REPETIDO

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.03	8	10	1500	25549.41	0.8956827	15542.42	5437.665	0.0442079	1614.973
0.03	8	10	1000	25563.24	0.8955566	15511.42	5478.169	0.0445978	1610.460
0.03	5	10	1500	25563.80	0.8951847	15492.91	5683.876	0.0482175	1611.261
0.03	5	10	1000	25625.86	0.8946270	15538.73	5732.542	0.0489431	1603.912
0.02	5	10	1500	25646.25	0.8944363	15557.22	5704.657	0.0480680	1535.341
0.02	8	10	1500	25701.18	0.8942308	15508.87	5699.058	0.0465466	1710.253

Após a redução de variáveis os modelos que se destacaram foram os que apresentaram os seguintes parâmetros:

BOOTSTRAP

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.03	5	10	1500	28120.75	0.8766462	17099.12	3561.382	0.0352837	912.4151
0.03	5	10	1000	28167.85	0.8762339	17099.26	3542.667	0.0350167	905.4760
0.02	5	10	1500	28169.29	0.8763379	17091.76	3344.293	0.0324422	908.0753
0.02	5	10	1000	28348.97	0.8747764	17173.14	3333.763	0.0324524	896.4338
0.03	3	10	1500	28354.08	0.8750307	17299.00	3523.955	0.0338864	839.9492
0.02	8	10	1500	28391.08	0.8742034	17107.80	3510.531	0.0347056	991.5874

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.03	5	10	1500	28120.75	0.8766462	17099.12	3561.382	0.0352837	912.4151
0.03	5	10	1000	28167.85	0.8762339	17099.26	3542.667	0.0350167	905.4760
0.02	5	10	1500	28169.29	0.8763379	17091.76	3344.293	0.0324422	908.0753
0.02	5	10	1000	28348.97	0.8747764	17173.14	3333.763	0.0324524	896.4338
0.03	3	10	1500	28354.08	0.8750307	17299.00	3523.955	0.0338864	839.9492
0.02	8	10	1500	28391.08	0.8742034	17107.80	3510.531	0.0347056	991.5874

K-FOLD REPETIDO

BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.03	5	10	1500	28120.75	0.8766462	17099.12	3561.382	0.0352837	912.4151
0.03	5	10	1000	28167.85	0.8762339	17099.26	3542.667	0.0350167	905.4760
0.02	5	10	1500	28169.29	0.8763379	17091.76	3344.293	0.0324422	908.0753
0.02	5	10	1000	28348.97	0.8747764	17173.14	3333.763	0.0324524	896.4338
0.03	3	10	1500	28354.08	0.8750307	17299.00	3523.955	0.0338864	839.9492
0.02	8	10	1500	28391.08	0.8742034	17107.80	3510.531	0.0347056	991.5874

K-FOLD REPETIDO

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.02	5	10	1500	26317.23	0.8892473	16360.57	5059.104	0.0440172	1556.678
0.02	8	10	1500	26323.56	0.8894248	16346.46	4881.701	0.0410800	1521.269
0.02	8	10	1000	26342.53	0.8891406	16326.37	4994.416	0.0424781	1522.205
0.01	8	10	1500	26375.67	0.8892154	16345.50	4999.524	0.0422016	1525.787
0.03	8	10	1000	26392.56	0.8891403	16477.80	4854.114	0.0408751	1530.685
0.03	5	10	1500	26420.00	0.8885045	16429.42	5015.108	0.0435132	1549.666

Portanto se decidiu utilizar o modelo que apresentou os melhores resultados utilizando K-Fold repetido e com menos variáveis, a tabela abaixo mostra quais foram os hiperparâmetros utilizados:

Portanto se decidiu utilizar o modelo que apresentou os melhores resultados utilizando K-Fold repetido e com menos variáveis, a tabela abaixo mostra quais foram os hiperparâmetros utilizados:

shrinkage	0.02
interaction.depth	5
n.minobsinnode	10
n.trees	1500
RMSE	26317.23
R²	0.8892473
MAE	16360.57

Realizando a determinação do contratante de que 80% da base deveria ser utilizada para treinar o modelo. Utilizando os hiperparâmetros selecionados, foi construído o modelo final que obteve os seguintes resultados na amostra teste.

Realizando a determinação do contratante de que 80% da base deveria ser utilizada para treinar o modelo. Utilizando os hiperparâmetros selecionados, foi construído o modelo final que obteve os seguintes resultados na amostra teste.

RMSE	36429.08
<hr/>	
R²	0.86
<hr/>	
MAE	19152.71

CONCLUSÃO

Foi obtido um R^2 de 86%, isto é, 86% da variação dos dados é explicada pelo modelo. O RMSE é alto pois a variância dos preços das casas é muito alta, por exemplo temos casas precificadas de 34900 até 755000, com desvio padrão de 79442.

Apesar do modelo K-fold Repetido ter demorado mais, ele apresentou melhores modelos comparados com os de bootstrap, e como a redução de variáveis diminuiu consideravelmente o tempo de treinamento ao custo de pouco desempenho optou-se por prosseguir com este modelo.

OBRIGADO!