

Nanodegree Engenheiro de Machine Learning

**Implementação de um modelo preditivo para preços de apartamentos
em Porto Alegre utilizando aprendizado de máquina**

Max Cohen

27/Jan/2019

SUMÁRIO

Listas de Figuras	3
Listas de Gráficos.....	3
Listas de Tabelas.....	3
1. Introdução.....	4
2. Metodologia.....	5
2.1. Conjunto de dados e entradas	5
2.2. Pré-processamento dos dados	6
2.3. Técnicas e Algoritmos.....	7
3. Análise Exploratória dos Dados	11
4. Resultados	27
5. Conclusão	31
5.1. Resultados obtidos	31
5.2. Aperfeiçoamento do modelo.....	31
6. Referências Bibliográficas	32

[Lista de Figuras](#)

Figura 1 – Tela com a descrição de um imóvel da Imobiliária Foxter.....	5
Figura 2 – Mapa	11

[Lista de Gráficos](#)

Gráfico 1 – Quantidade de imóveis por bairro.....	12
Gráfico 2 – Histogramas das variáveis área, box, valor condomínio, valor do IPTU, quantidade de quartos e preço.....	13
Gráfico 3 – Boxplot do preço de venda por bairro.....	15
Gráfico 4 – Quantidade de Quartos por Bairro	16
Gráfico 5 – Quantidade de Apartamentos por Quantidade de Quartos e Bairro.....	16
Gráfico 6 – Garagens por Bairro	17
Gráfico 7 – Bairros por Garagens.....	17
Gráfico 8 – Boxplot da área (m^2)	18
Gráfico 9 – Boxplot do valor do condomínio	19
Gráfico 10 – Boxplot do valor do IPTU	20
Gráfico 11 – Pairplot: relacionamento entre as variáveis	21
Gráfico 12 – Correlações	22
Gráfico 13 – Regressão linear entre as variáveis área e preço	23
Gráfico 14 – Regressão linear entre as variáveis área e preço do m^2	23
Gráfico 15 – Pairplot: relacionamento entre todas as variáveis destacando os bairros	24
Gráfico 16 – Distribuição dos imóveis por área x preço x bairro	25
Gráfico 17 – Distribuição dos imóveis por área x preço do m^2 x bairro	25
Gráfico 18 – Regressão linear entre as várias área e preço com agrupamento por bairro.....	26
Gráfico 19 – Comparação dos Algoritmos	28
Gráfico 20 – Importância dos Atributos	29
Gráfico 21 – Importância Acumulada	30

[Lista de Tabelas](#)

Tabela 1 – Estatísticas descritivas do preço, área, condomínio, IPTU, quartos e garagem....	14
Tabela 2 – Médias das variáveis preço, área, condomínio, IPTU, quartos e garagem identificados por bairro.....	14
Tabela 3 – Quantidade de apartamentos por quantidade de quartos e bairro	16
Tabela 4 – Estatísticas descritivas para a quantidade de quartos por bairro.....	16
Tabela 5 – Quantidade de apartamentos por quantidade de garagens e bairro	17
Tabela 6 – Estatísticas descritivas para a quantidade de garagens por bairro.....	17

1. Introdução

A economia brasileira entrou em crise há alguns anos e continua sem dar sinais de melhoria para o curto prazo. No período entre 2012 e 2013 os preços dos imóveis alcançaram a maior alta histórica. Naquele momento o mercado estava aquecido, com uma forte demanda diante da oferta. Hoje o cenário é bem diferente.

Depois do *boom* os preços começaram a cair de forma constante e, até o momento, ininterrupta. Diante da deterioração da economia, muitos brasileiros começaram a vender seus imóveis, ou para saldar dívidas ou para poder deixar o país. Dia após dia os preços dos imóveis eram não só diferentes mas também menores. Eu me vi nessa situação quando passei a ofertar a venda do meu apartamento e fui obrigado a reduzir o preço de forma sistemática. Contudo, naquele momento, uma dúvida permanecia: “quanto vale o meu apartamento no atual momento do mercado?”

Na literatura acadêmica, assim como em trabalhos técnicos diversos, é possível encontrar muitos artigos, teses e relatórios sobre a aplicação de aprendizado de máquina para previsão de preços imóveis (BERNÁDEZ, 2018; KOMAGOME-TOWNE, 2016; NGUYEN, 2018; PARK & BAE, 2018; PIERRE, 2018). É possível encontrar também trabalhos utilizando redes neurais (*deep learning*) (YU, 2018). O maior benefício no uso do aprendizado de máquina para resolução desse tipo de problema é a possibilidade do uso de vários algoritmos e a escolha do que apresenta melhor desempenho. Em outras palavras, é poder ter um algoritmo otimizado para o problema proposto.

Tendo em vista a piora da economia e a constante queda dos preços dos apartamentos usados, tem-se uma dificuldade de se identificar, de forma científica e rápida, qual é o preço justo para um apartamento. O alvo deste estudo foi a cidade de Porto Alegre (RS). Desta forma, o problema principal desta pesquisa foi: *Qual é o preço de venda de um determinado apartamento usado na cidade de Porto Alegre?*

Para direcionar o trabalho, elencou-se as seguintes perguntas secundárias:

- 1) Qual é o melhor algoritmo para a amostra estudada?
- 2) Quais os atributos que mais influenciam no preço de venda?

O “preço justo de venda” aqui declarado se refere ao preço praticado num determinado momento e compatível com a realidade do mercado. O “momento” para esta pesquisa foi determinado com um dia específico. Por sua vez, o preço de um apartamento pode ser estimado a partir das características do imóvel, como tamanho, quantidade de dormitórios, vagas de garagem, etc. Entende-se, portanto, que se está tratando de um problema supervisionado, de regressão, cujo “preço de venda” será a variável dependente, com seu valor expresso em unidades monetárias.

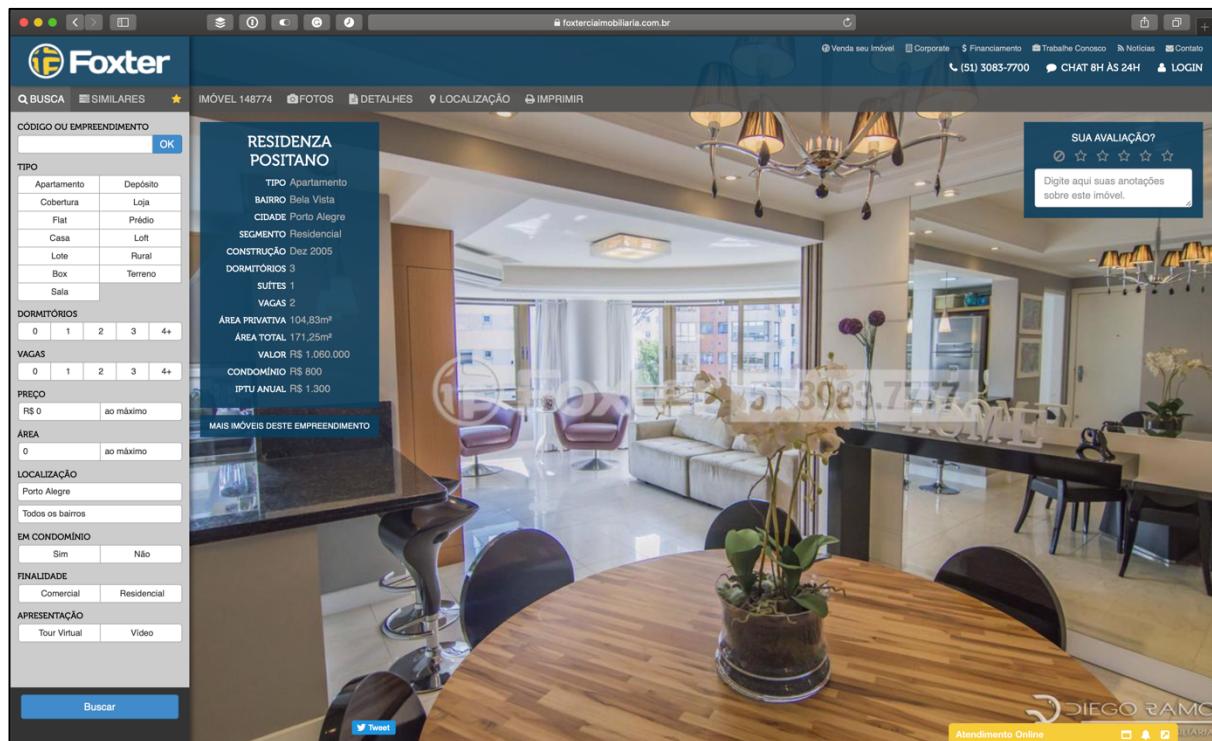
2. Metodologia

O projeto foi desenvolvido em três etapas. Na primeira etapa foi construído um *web scraper* que capturou os dados disponíveis sobre os imóveis no site da imobiliária Foxter e armazenou num arquivo *csv*. A segunda etapa constituiu na leitura do *csv*, limpeza e formação dos dados. A terceira etapa foi a análise exploratória dos dados. E, por último, a quarta etapa consistiu na criação do modelo e avaliação dos algoritmos.

2.1. Conjunto de dados e entradas

Nos últimos anos, pequenas e grandes imobiliárias em Porto Alegre têm atuado na venda de imóveis novos e usados, tendo seus portfolios disponibilizados de forma pública na Internet. Diante disso, o conjunto de dados trabalhado neste projeto foi composto pelos dados disponíveis no site da imobiliária Foxter.

Figura 1 – Tela com a descrição de um imóvel da Imobiliária Foxter¹



Um *web scraper* foi criado especificamente com o objetivo de extrair os dados descritivos dos apartamentos disponíveis e povoados em um *dataset*. Os dados dos imóveis estão disponíveis em um retângulo de fundo azul na tela de consulta da imobiliária, como no exemplo da figura acima. Assim, os dados coletados foram:

¹ <https://www.foxterciamobiliaria.com.br/imovel/148774/residencial-porto-alegre-bela-vista-apartamento-residenza-positano-3-dormitorios-zona-norte>

1. De identificação:
 - 1.1. código do imóvel;
 - 1.2. url;
2. Target:
 - 2.1. preço;
3. Atributos:
 - 3.1. área;
 - 3.2. bairro;
 - 3.3. cidade;
 - 3.4. tipo;
 - 3.5. segmento;
 - 3.6. valor do condomínio;
 - 3.7. valor do IPTU;
 - 3.8. quantidade de dormitórios;
 - 3.9. quantidade de vagas de estacionamento.

A quantidade de imóveis ofertados pelo site deve variar diariamente por causa das vendas e da inserção de novas ofertas. Os dados utilizados neste estudo foram coletados pelo *web scraper* no dia 20/01/2019. O estudo foi delimitado aos dados dos bairros Auxiliadora, Bela Vista e Mont’Serrat, por serem vizinhos e manterem semelhanças entre si. Desta forma, o aplicativo do *web scraper* foi ajustado incluindo essa delimitação. Foram coletados dados de 577 imóveis.

2.2. Pré-processamento dos dados

Foi executado uma pré-processamento dos dados antes da análise exploratória com o objetivo de checagem desses dados. A primeira ação foi a leitura do arquivo csv, criado pelo *web scraper*, e armazenamento dos dados num Pandas Data Frame. Totalizou-se 577 imóveis. Em seguida checou-se se havia ou não imóveis duplicados a partir do seu código de identificação. Não foram encontrados imóveis repetidos. Eliminou-se do data frame as colunas desnecessárias, 'Unnamed: 0' e 'id'.

Na visualização dos atributos por bairros, foi identificado que alguns imóveis não tinham preço. Esse era um indicativo que havia valores faltantes no conjunto. Outros bairros, além dos três delimitados pelo trabalho, apareceram e foram retirados do data frame. Checou-se também se os imóveis eram realmente de Porto Alegre (cidade), apartamento (tipo), residencial (segmento). O atributo “rooms” que armazenava a quantidade de quartos apresentou um único valor estranho que foi retirado.

Em seguida foram eliminados do data frame as colunas que não fariam parte da análise: 'city', 'type', 'segment', 'url', 'date'. Assim como valores discrepantes (outliers) que foram detectados. Por fim, o data frame contabilizou o total de 287 imóveis.

2.3. Técnicas e Algoritmos

Para alcançar o objetivo do trabalho em estimar o preço de venda de um apartamento em Porto Alegre, buscou-se a aplicação de um algoritmo regressor, cujo conjunto de atributos levados em consideração foram:

1. Área do apartamento ('area')
2. Valor do condomínio ('condominium')
3. Valor do IPTU ('iptu')
4. Quantidade de vagas de garagem ('box')
5. Quantidade de quartos ('rooms')
6. Bairro ('district')

A análise foi focada em três bairros: Auxiliadora, Bela Vista e Mont’Serrat. Desta forma, transformou-se a variável de identificação do bairro em duas variáveis *dummies*, passando a serem identificadas como: 'district_Bela Vista' e 'district_Mont Serrat'. A variável original 'district' foi retirada do *data frame*, visto que não era mais necessária.

Para avaliação dos modelos a serem gerados, três condições foram determinadas inicialmente:

1. A estratégia de divisão dos dados para a validação cruzada foi determinada para se usar 10 conjuntos²;
2. Foi escolhido o método KFold³ como estratégia para a validação cruzada.
3. Um número randômico foi estabelecido como 42, de forma que o trabalho possa ser repetido e comparável;

Os dados originais foram divididos em quatro conjuntos, sendo dois para treino (X_train, y_train) e dois para teste (X_test, y_test). O atributo preço do *data frame* passou a integrar o “y” e os demais atributos o “X”. A divisão foi determinada para que 25% do total dos dados fossem para teste e o restante para treino.

O algoritmo de regressão linear⁴ foi selecionado como modelo de benchmark para o trabalho. A regressão linear é uma técnica clássica no ferramental estatístico, usada amplamente para determinação de variáveis dependentes contínuas.

Um conjunto de algoritmos regressores foi criado com intuito de se identificar qual teria o melhor desempenho. *A priori* não se tinha nenhuma informação de qual seria o melhor para o caso em estudo. Partiu-se, portanto, de uma estratégia de seleção aleatória⁵, onde foram elencados e testados oito algoritmos. A estratégia teve como benefícios⁶ a velocidade obtida para o trato com diferentes algoritmos; objetividade na aplicação de diferentes algoritmos para um único problema definido; e a obtenção de resultados comparáveis. Os algoritmos foram:

² cv=10, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

³ https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

⁴ https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

⁵ Estratégia “spot-checking” (p.76, BROWNLEE, 2018).

⁶ Da mesma forma como relatado por Brownlee (2014).

1. Linear Regression⁷
2. Lasso⁸
3. Elastic Net⁹
4. K-Neighbors Regressor¹⁰
5. Decision Tree Regressor¹¹
6. Support Vector Regression¹²
7. AdaBoost Regressor¹³
8. Gradient Boosting Regressor¹⁴

A regressão linear (*linear regression*) é uma técnica estatística consagrada e amplamente utilizada. Ela descreve o relacionamento de variáveis, onde o modelo é construído com coeficientes para minimizar o erro pela diferença do conjunto de observações e os valores previstos pela aproximação linear. A existência de variáveis com alta correlação leva à situação de multicolinearidade¹⁵.

A regressão Lasso (*Least Absolute Shrinkage and Selection Operator*), por sua vez, é uma variação da regressão linear, onde a função *loss* é modificada para minimizar a complexidade do modelo, reduzindo o número de variáveis¹⁶.

A regressão *Elastic Net* combina dois tipos de regressores: Ridge e Lasso. Busca minimizar a complexidade do modelo a partir dos coeficientes L2 (a soma dos quadrados dos valores dos coeficientes) e L1 (a soma absoluta dos valores dos coeficientes). É útil quando há vários atributos correlacionados entre si¹⁷.

O K-Neighbors (k vizinhos), atua de forma não linear, localizando as k instâncias similares no conjunto de treinamento. Trata-se de algoritmo clássico, simples e de fácil entendimento¹⁸. Dos k vizinhos a média ou mediana das variáveis é gerado e se torna o preditor.

CART, ou *Decision Tree Regressor*, é um algoritmo de aprendizado supervisionado não-paramétrico usado tanto para classificação como para regressão. Usa o conjunto de dados para selecionar os melhores pontos, dividindo os dados de forma a minimizar o custo da métrica de desempenho (que usualmente é o erro quadrado médio – MAE). Como vantagem tem-se: seu uso e interpretação são simples; as árvores criadas podem ser visualizadas; requer pouca preparação dos dados; capaz de lidar com dados numéricos e categóricos; é capaz de lidar com problemas de múltiplas saídas¹⁹.

SVR, ou *Support Vector Regression*, tem como objetivo encontrar uma função que apresente o menor erro possível determinado por um intervalo. Essa função busca separar os

⁷ https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

⁸ https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

⁹ https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

¹⁰ <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>

¹¹ <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>

¹² <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

¹³ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html>

¹⁴ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>

¹⁵ https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

¹⁶ https://scikit-learn.org/stable/modules/linear_model.html#lasso

¹⁷ https://scikit-learn.org/stable/modules/linear_model.html#elastic-net

¹⁸ <https://bit.ly/2Ur4aSc>

¹⁹ <https://scikit-learn.org/stable/modules/tree.html#tree>

dados por um hiperplano. A otimização da solução passa por maximizar a margem divisória, ou seja, conseguir o maior espaçamento possível de dois conjuntos de dados distintos separados pelo hiperplano (BHATTACHARYYA, 2018). No caso do SVR, o erro é contabilizado pela soma dos erros de classificação e dos erros de margem²⁰. No caso dos dados que não são linearmente separáveis no espaço dimensional original, esses podem ser linearmente separáveis em um espaço dimensional mais alto. Essa abordagem é denominada de “truque do kernel”²¹ (KANDAN, 2017).

AdaBoost (ou *Adaptative Boosting*) é um algoritmo que faz parte de um método denominado *Boosting Ensemble* (conjunto impulsor) que cria um classificador “forte” a partir de um número de classificadores “fracos”. Inicialmente um modelo é treinado, sendo seguido por segundo modelo que corrige os erros do primeiro. Os modelos são adicionados até se ter um conjunto perfeito ou limitado por um número máximo de modelos adicionados. AdaBoost foi criado inicialmente para classificação e depois para atuar como regressor (BROWNLEE, 2016).

Gradient Boosting Regressor é um algoritmo que constrói um modelo aditivo de maneira progressiva, ou seja, em cada estágio uma árvore de regressão é ajustada²². Trata-se de uma generalização de estímulo/impulso para funções de perda arbitrariamente diferenciáveis²³. É utilizado tanto para classificação como para regressão. Tem como vantagens o trato de dados mistos (heterogêneos); força preditiva; robustez para tratar dados discrepantes.

Por se tratar de uma regressão, a avaliação do desempenho do modelo gerado foi comparado com um modelo prévio para *benchmark* e teve computado as métricas específicas usualmente empregadas (p.299, Müller & Guido, 2017) (Swalin, 2018), sendo: MAE²⁴ (*Mean Absolute Error*), MSE²⁵ (*Mean Squared Error*) e R² (também chamado de Coeficiente de Determinação)²⁶. As equações das métrica são^{27,28,29}:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|.$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

²⁰ <https://bit.ly/2UrmBWL>

²¹ <https://bit.ly/2FW4jTH>

²² <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>

²³ <https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting>

²⁴ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

²⁵ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

²⁶ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

²⁷ https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error

²⁸ https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error

²⁹ https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1} (y_i - \bar{y})^2}$$

O algoritmo com o melhor desempenho da métrica R² foi selecionado e teve seus parâmetros ajustados em busca de um refinamento e de um desempenho ainda superior. Para tanto, utilizou do GridSearchCV³⁰ que, a partir de uma configuração inicial, executou uma busca exaustiva para encontrar uma combinação de parâmetros que potencialize o desempenho do algoritmo testado.

A motivação pela escolha da estatística R² foi por fornecer “(...) uma medida de quão bem as amostras futuras provavelmente serão previstas pelo modelo”³¹.

³⁰ https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

³¹ https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

3. Análise Exploratória dos Dados

Os dados utilizados no estudo foram coletados do site da imobiliária Foxter³², uma das maiores de Porto Alegre. Com o intuito de delimitar a pesquisa, foram escolhidos os bairros Auxiliadora, Bela Vista e Mont’Serrat como alvos, devido serem vizinhos (ver Figura 2), partilhando de ruas e linhas de ônibus em comum, e com infraestrutura (supermercados, farmácias etc.) semelhante.

Figura 2 – Mapa³³

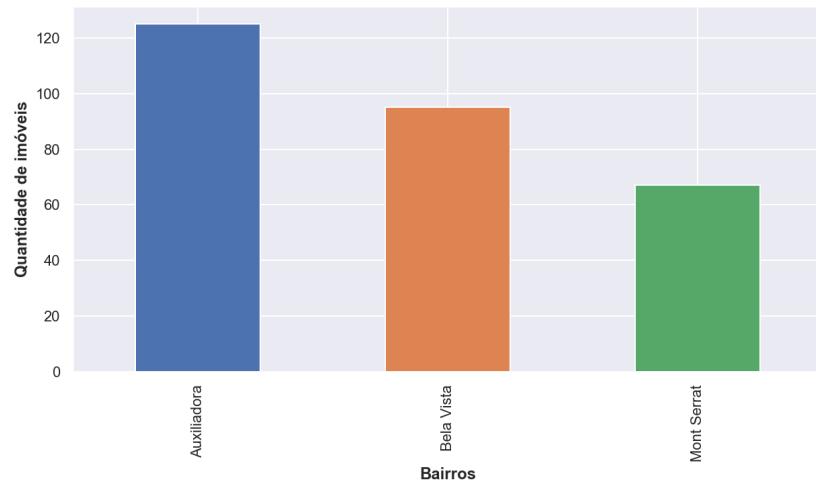


³² <https://www.foxterciaimobiliaria.com.br>

³³ Fonte: Google Maps.

A amostra totalizou 287 imóveis. O Gráfico 1 apresenta a quantidade de imóveis por bairro, sendo: 125 do bairro Auxiliadora, 95 da Bela Vista e 67 de Mont’Serrat.

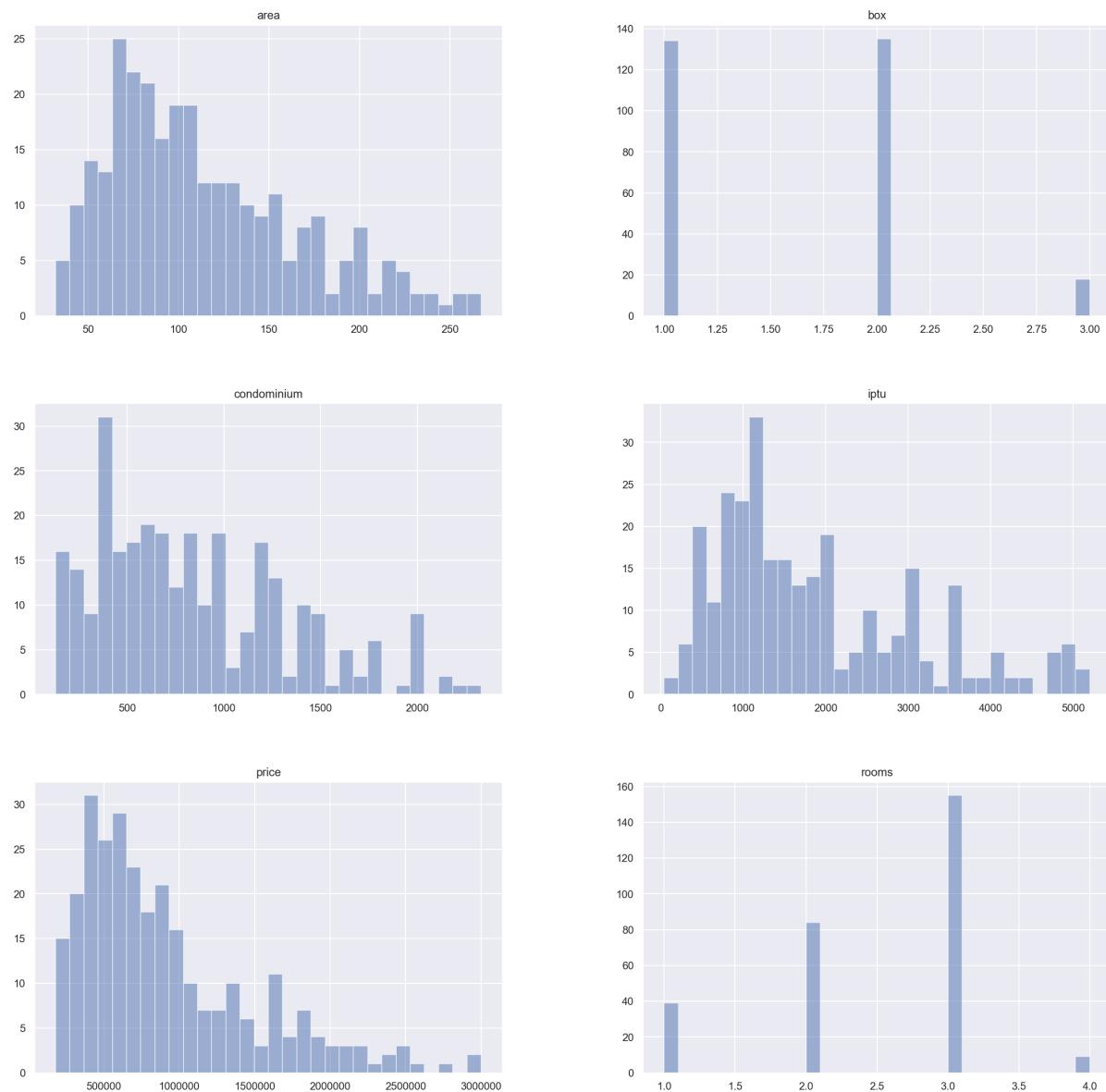
Gráfico 1 – Quantidade de imóveis por bairro



O Gráfico 2 apresenta os histogramas das seis variáveis quantitativas. Nenhuma delas teve comportamento igual ou próximo da curva normal. As variáveis indicativas de valores monetários, como condomínio, IPTU e preço do imóvel, assim como a variável que registrou a área em metros quadrados, apresentaram desvios à direita. Isso já seria esperado, visto que os maiores apartamentos são aqueles que possuem maiores preços de venda, de IPTU e de condômino, e ao mesmo tempo estão em quantidade menores.

Também é possível observar que no quesito vagas de garagem, a variável “box” registrou que a quantidade de apartamentos com uma vaga é praticamente igual aos de duas vagas. Apartamento com três vagas são poucos. E quanto a quantidade de quartos, apartamentos com três quartos formam a maioria dos anúncios.

Gráfico 2 – Histogramas das variáveis área, box, valor condomínio, valor do IPTU, quantidade de quartos e preço



Na tabela a seguir estão as estatísticas descritivas das variáveis quantitativas. Há uma grande variação dos preços, sendo o mínimo de R\$180 mil e o máximo R\$3 milhões. O valor do condomínio e IPTU também apresentam grandes variações, com valores entre R\$130 e R\$2,3 mil, e de R\$43 a R\$5,2 mil, respectivamente.

Tabela 1 – Estatísticas descritivas do preço, área, condomínio, IPTU, quartos e garagem

	Preço	Área	Valor do Condomínio	IPTU	Quant. de Quartos	Vagas de Garagem
quantidade	287.00	287.00	287.00	287.00	287.00	287.00
média	913,783.59	113.66	842.99	1,875.40	2.47	1.60
std	579,992.70	52.32	501.85	1,215.94	0.77	0.61
min	180,000.00	32.00	130.00	43.00	1.00	1.00
25%	480,000.00	72.88	440.00	995.00	2.00	1.00
50%	735,000.00	101.72	750.00	1,500.00	3.00	2.00
75%	1,200,000.00	144.26	1,200.00	2,624.00	3.00	2.00
max	3,001,000.00	267.34	2,330.00	5,200.00	4.00	3.00

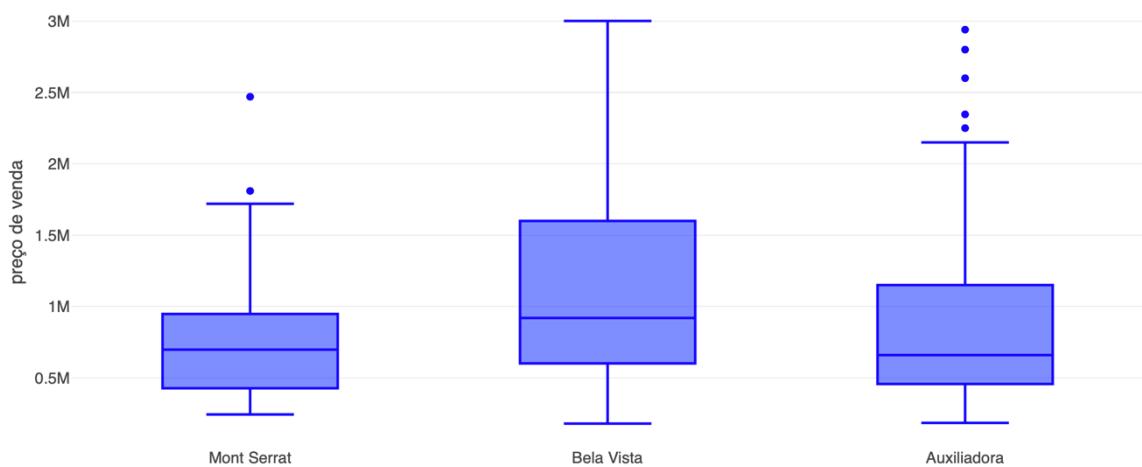
Na tabela seguinte estão as médias das variáveis quantitativas. É possível observar que o bairro Bela Vista apresenta as maiores médias. É, portanto, o bairro com os maiores e mais caros apartamentos dentre os três analisados.

Tabela 2 – Médias das variáveis preço, área, condomínio, IPTU, quartos e garagem identificados por bairro

Bairros	Preço	Área	Valor do Condomínio	IPTU	Quant. de Quartos	Vagas de Garagem
Auxiliadora	859,152.22	110.88	796.66	1,778.84	2.49	1.60
Bela Vista	1,090,082.76	125.94	937.54	2,149.37	2.51	1.62
Mont'Serrat	765,731.34	101.43	795.34	1,667.09	2.37	1.55

No tocante à preços, o Bairro da Bela Vista apresenta o maior preço médio do conjunto estudado. No boxplot a seguir, é possível observar que tanto Mont’Serrat quanto Auxiliadora apresentam imóveis com valores altos semelhantes à Bela Vista. Contudo, o mesmo boxplot chama atenção quando identifica tais valores como discrepantes ao seu conjunto original do bairro.

Gráfico 3 – Boxplot do preço de venda por bairro



Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	859,152.22	578,902.68	185,000.00	460,000.00	660,000.00	1,150,000.00	2,940,000.00
Bela Vista	95.00	1,090,082.76	632,531.82	180,000.00	604,500.00	920,000.00	1,599,500.00	3,001,000.00
Mont Serrat	67.00	765,731.34	430,795.34	244,000.00	435,000.00	698,000.00	947,000.00	2,470,000.00

Os Gráfico 4, Gráfico 5, Tabela 4 e Tabela 4 descrevem a quantidade de quartos dos imóveis analisados. É possível observar no gráfico que os apartamentos com três quartos são os de maior quantidade nos três bairros analisados. Apartamentos com quatro quartos são os de menores quantidade. São mais comuns em Bela Vista do que nos outros dois bairros.

Gráfico 4 – Quantidade de Quartos por Bairro

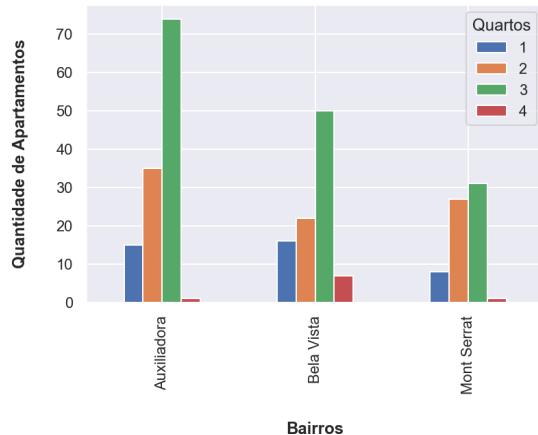


Gráfico 5 – Quantidade de Apartamentos por Quantidade de Quartos e Bairro

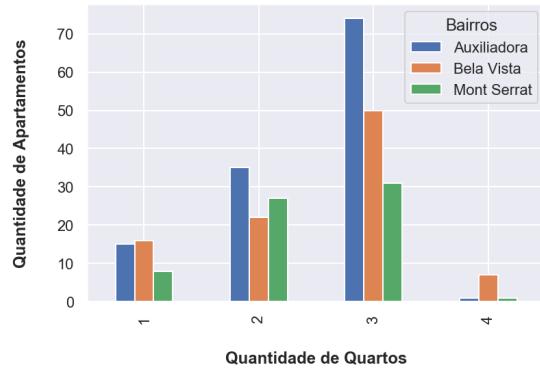


Tabela 3 – Quantidade de apartamentos por quantidade de quartos e bairro

Bairros		Auxiliadora		Bela Vista		Mont Serrat		Total	
Quartos									
1		15.00		16.00		8.00		39.00	
2		35.00		22.00		27.00		84.00	
3		74.00		50.00		31.00		155.00	
4		1.00		7.00		1.00		9.00	
Total		125.00		95.00		67.00		287.00	

Tabela 4 – Estatísticas descritivas para a quantidade de quartos por bairro

Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	2.49	0.71	1.00	2.00	3.00	3.00	4.00
Bela Vista	95.00	2.51	0.86	1.00	2.00	3.00	3.00	4.00
Mont'Serrat	67.00	2.37	0.71	1.00	2.00	2.00	3.00	4.00

Todos os apartamentos ofertados possuem ao menos uma vaga de garagem. No conjunto, 134 apartamentos possuem uma vaga, 135 possuem duas e 18 possuem três.

Gráfico 6 – Garagens por Bairro

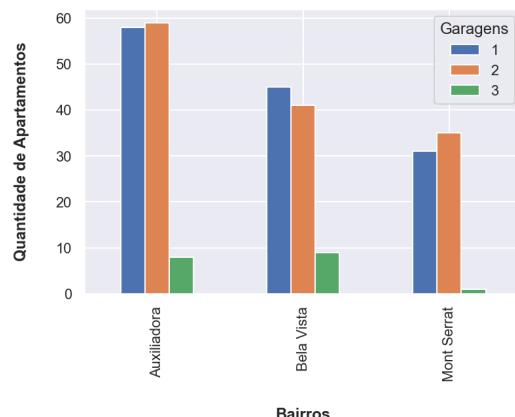


Gráfico 7 – Bairros por Garagens

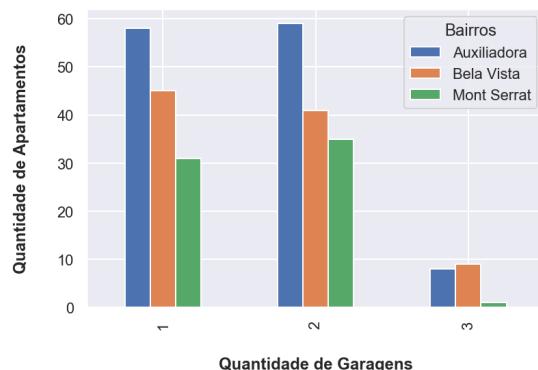


Tabela 5 – Quantidade de apartamentos por quantidade de garagens e bairro

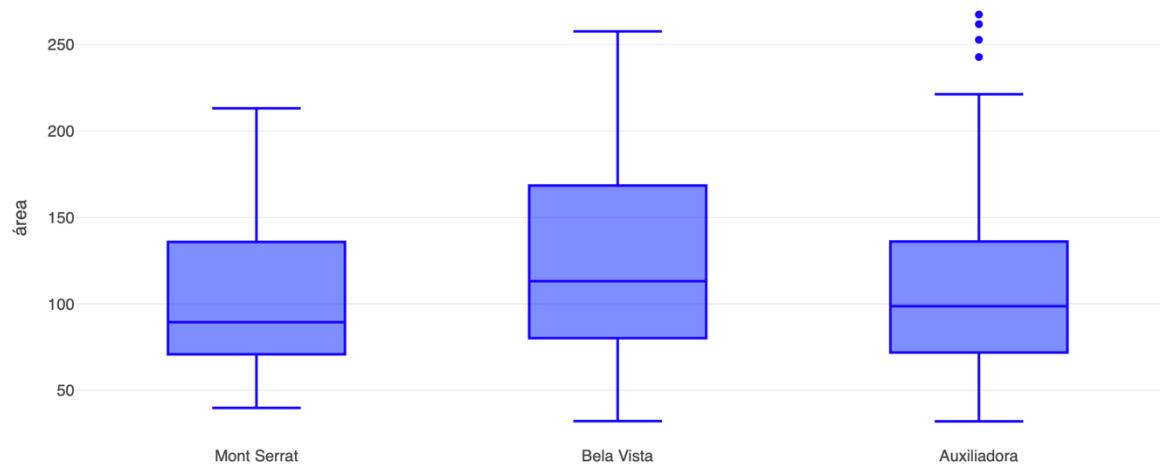
Bairros	Auxiliadora	Bela Vista	Mont Serrat	Total
Garagens				
1	58	45	31	134.00
2	59	41	35	135.00
3	8	9	1	18.00
Total	125.00	95.00	67.00	287.00

Tabela 6 – Estatísticas descritivas para a quantidade de garagens por bairro

Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	1.60	0.61	1.00	1.00	2.00	2.00	3.00
Bela Vista	95.00	1.62	0.66	1.00	1.00	2.00	2.00	3.00
Mont'Serrat	67.00	1.55	0.53	1.00	1.00	2.00	2.00	3.00

Visualizando o boxplot a seguir é possível verificar que a área dos apartamentos são semelhantes. Bela Vista possui a maior média para essa variável, com m^2 médio igual a 125,94. O bairro também possui apartamento pequenos (mínimo de 32,13), menores até que Mont'Serrat. Esse fato leva a apresentar o maior desvio-padrão dentre os três bairros.

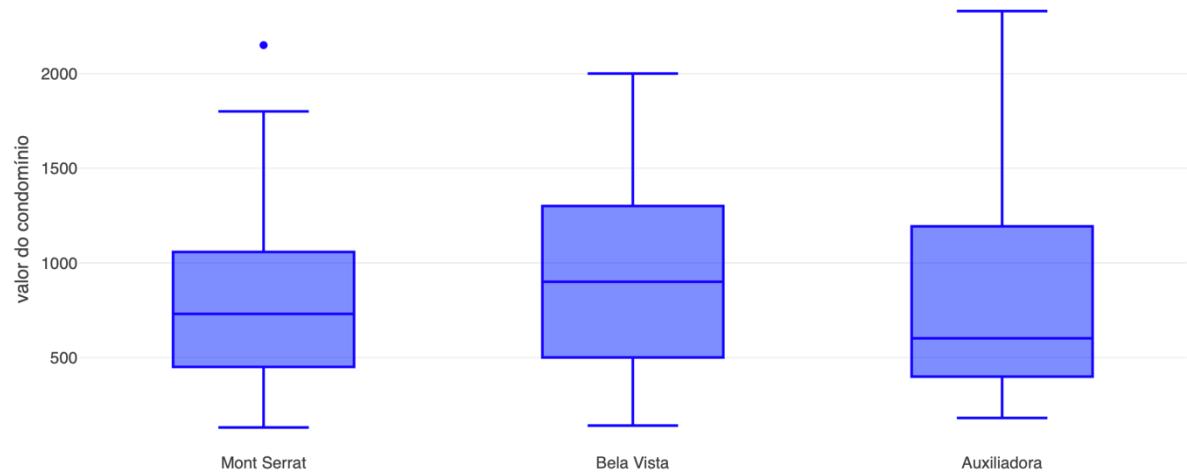
Gráfico 8 – Boxplot da área (m^2)



Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	110.88	51.28	32.00	72.09	98.61	136.00	267.34
Bela Vista	95.00	125.94	57.64	32.13	80.30	113.10	167.95	257.60
Mont'Serrat	67.00	101.43	42.49	39.77	70.85	89.38	134.72	213.10

O valor mínimo pago a título de condomínio foi identificado no bairro de Mont’Serrat, sendo de R\$130,00, seguido da Bela Vista com R\$140,00. O maior valor foi do bairro Auxiliadora, com R\$2330,00. Bela Vista possui o maior valor médio, R\$937,54.

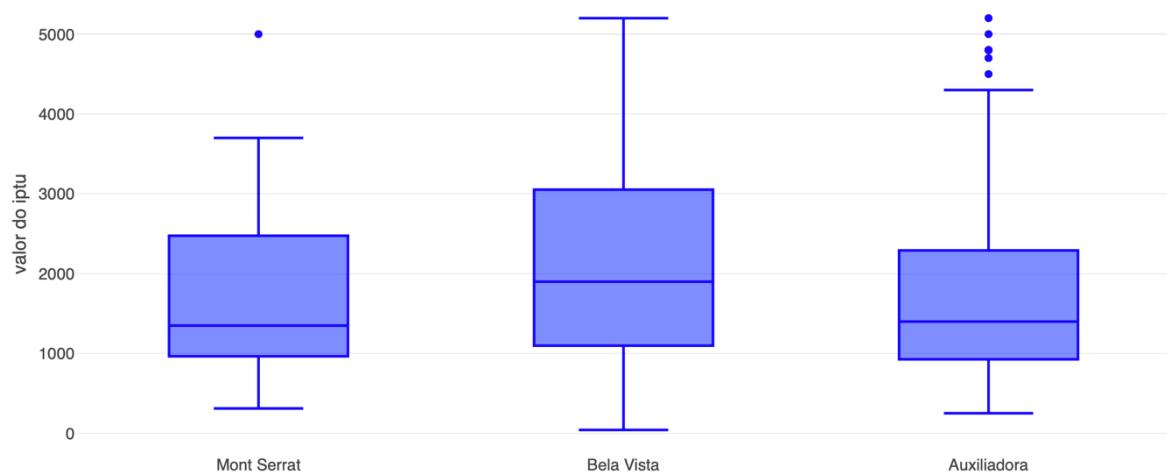
Gráfico 9 – Boxplot do valor do condomínio



Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	796.66	531.07	180.00	400.00	601.00	1190.00	2,330.00
Bela Vista	95.00	937.54	511.08	140.00	500.00	900.00	1300.00	2,000.00
Mont’Serrat	67.00	795.34	412.57	130.00	450.00	730.00	1051.50	2,150.00

O boxplot a seguir apresenta os valores do IPTU para os três bairros. Os três são semelhantes. Um ponto que chamou atenção foi o valor mínimo para Bela Vista, de R\$43,00. É um valor muito baixo para um bairro que tem apresentado indicativos de ser o bairro mais caro dos três analisados. Já os valores máximos são bem próximos. Bela Vista volta a apresentar o maior valor médio.

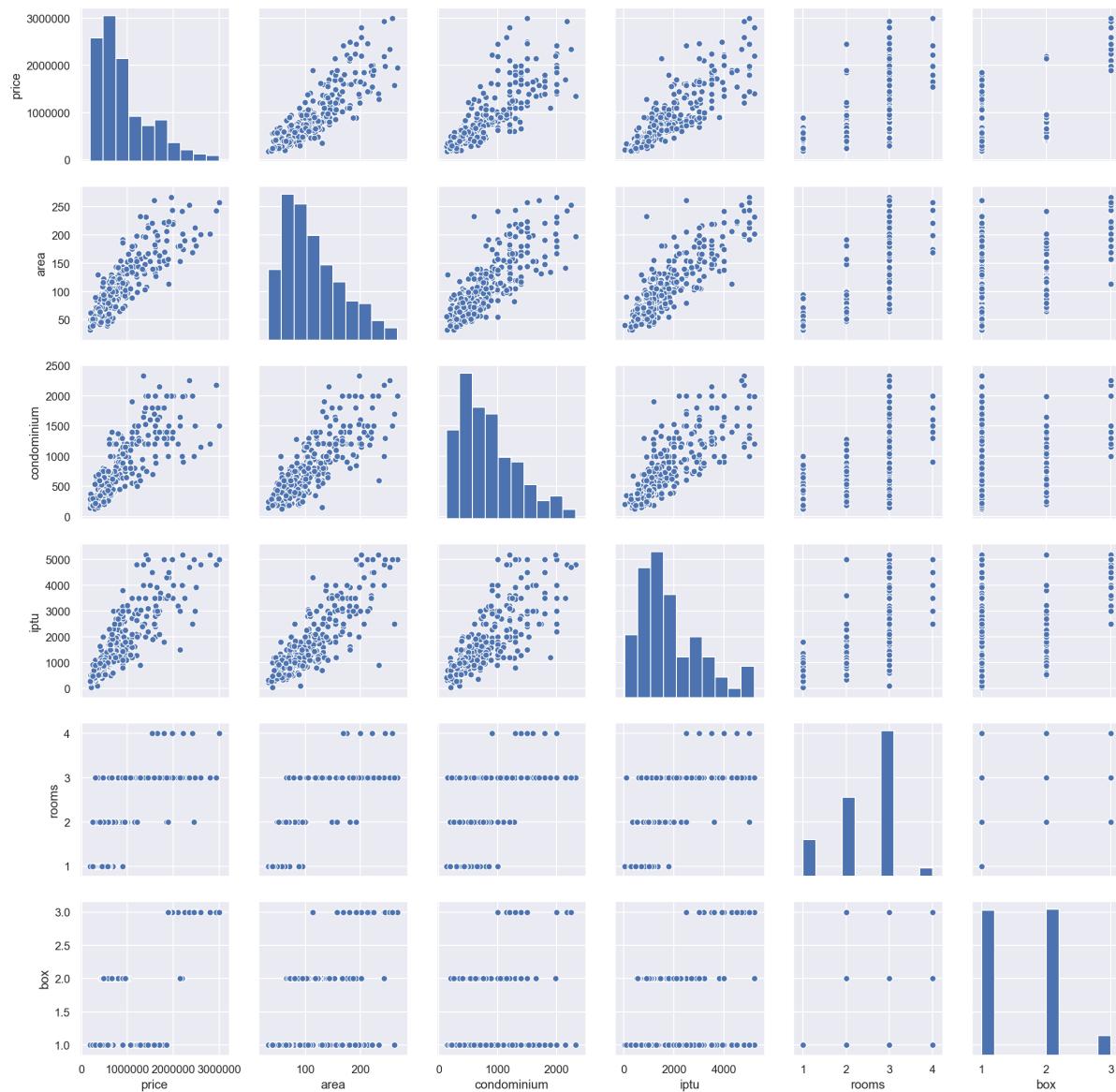
Gráfico 10 – Boxplot do valor do IPTU



Bairros	Quant.	Média	Desvio-Padrão	Min	25%	50%	75%	Max
Auxiliadora	125.00	1,778.84	1,194.30	252.00	930.00	1,400.00	2,288.00	5,200.00
Bela Vista	95.00	2,149.37	1,328.14	43.00	1,099.00	1,900.00	3,035.00	5,200.00
Mont Serrat	67.00	1,667.09	1,022.28	312.00	970.00	1,350.00	2,450.00	5,000.00

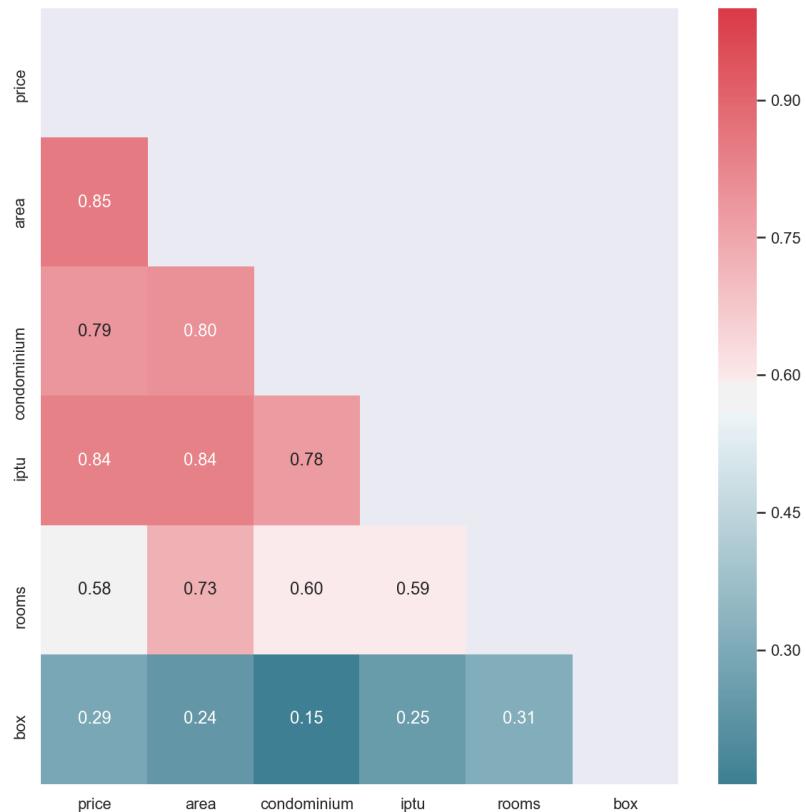
O *pairplot* abaixo apresenta os relacionamentos entre as variáveis a partir da dispersão gráfica dos pontos. É possível ver que o preço mantém uma relação positiva com todas as demais variáveis. As variáveis contínuas (área, valor do condomínio e IPTU) também apresentam relacionamentos positivos entre si.

Gráfico 11 – Pairplot: relacionamento entre as variáveis



O resultado das correlações permite quantificar o grau do relacionamento entre as variáveis e o sentido (positivamente ou negativamente). No gráfico seguinte tem-se os valores das correlações. A maioria resultou em valores medianos (entre 0,50 e 0,75) e fortes (acima de 0,750). Destaca-se a variável “box”, que congrega as quantidades de vagas de garagem, que teve baixa correlação (de 0,15 a 0,31) com todas as demais.

Gráfico 12 – Correlações



Com valor igual a 0,85, área e preço possuem uma alta correlação. No Gráfico 13 é possível observar tal relacionamento, cuja a reta gerada pela regressão linear reúne bem os pontos plotados. A reta demonstra uma inclinação próxima de 45 graus, indicando que os preços subirão em proporção igual à área. Verifica-se, portanto, como já esperado, que o aumento da área resulta no aumento do preço do imóvel.

No Gráfico 14 tem-se a relação da área com o preço do m². Nesse gráfico pode-se observar que apartamentos com a mesma área possuem valores diferentes de preço do m². Isso pode ocorrer quando compararmos imóveis novos com usados do mesmo tamanho, onde é comum que os imóveis novos possuam maior valoração m².

Gráfico 13 – Regressão linear entre as variáveis área e preço

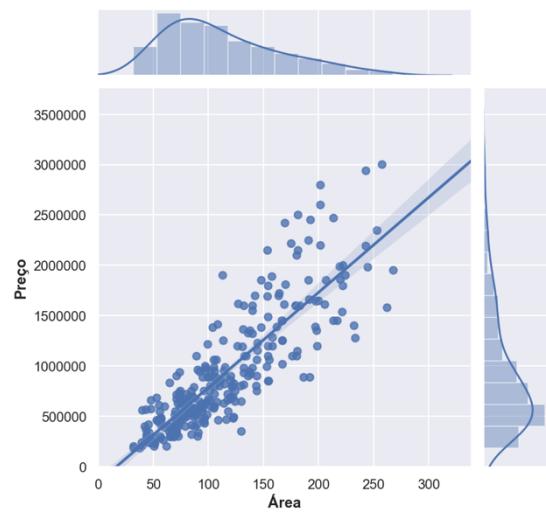
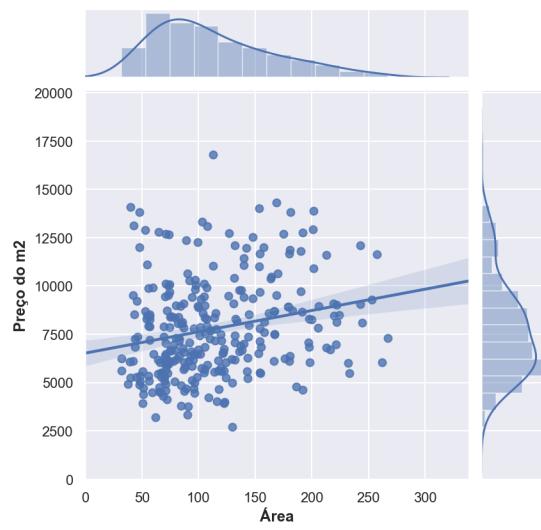
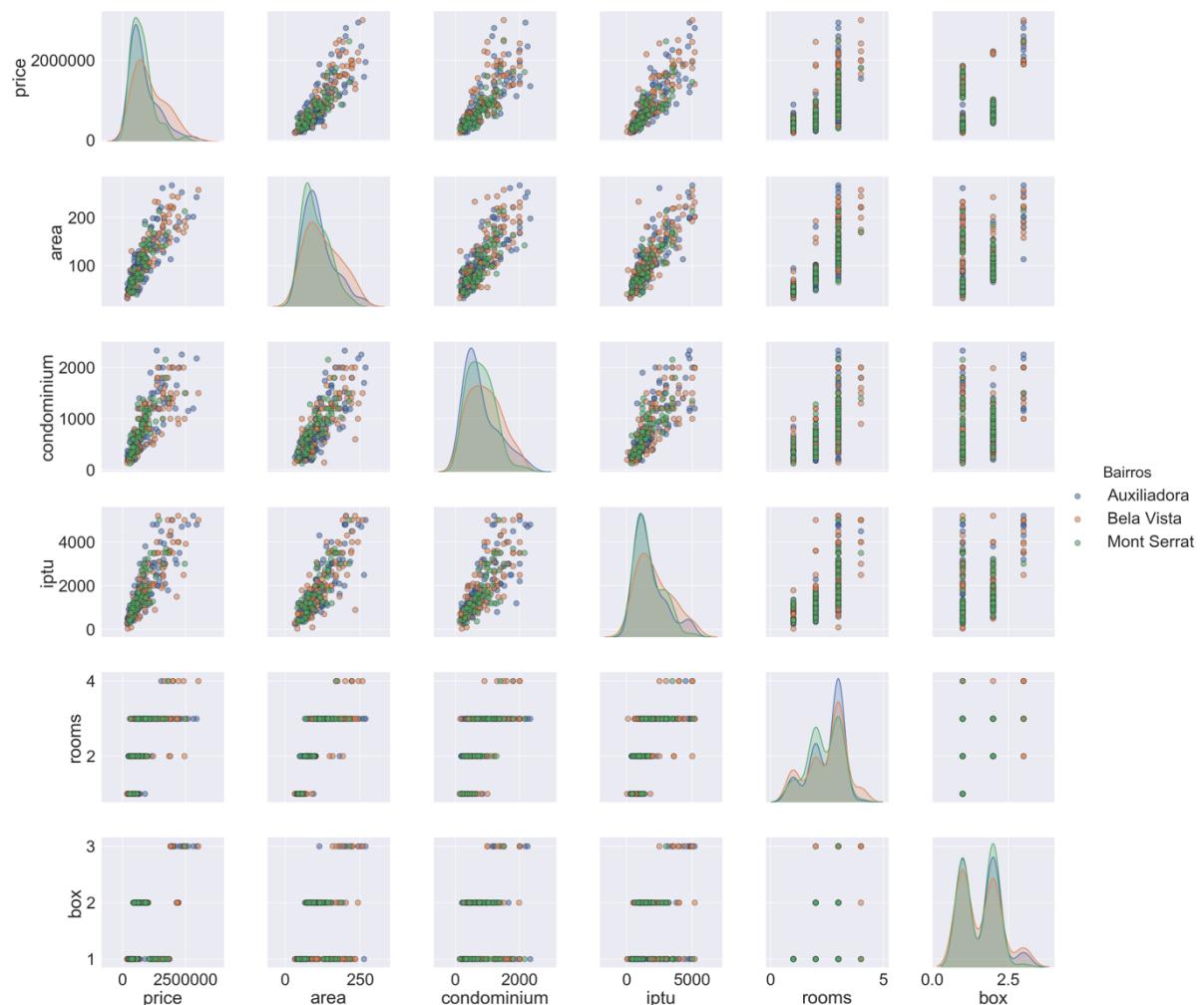


Gráfico 14 – Regressão linear entre as variáveis área e preço do m²



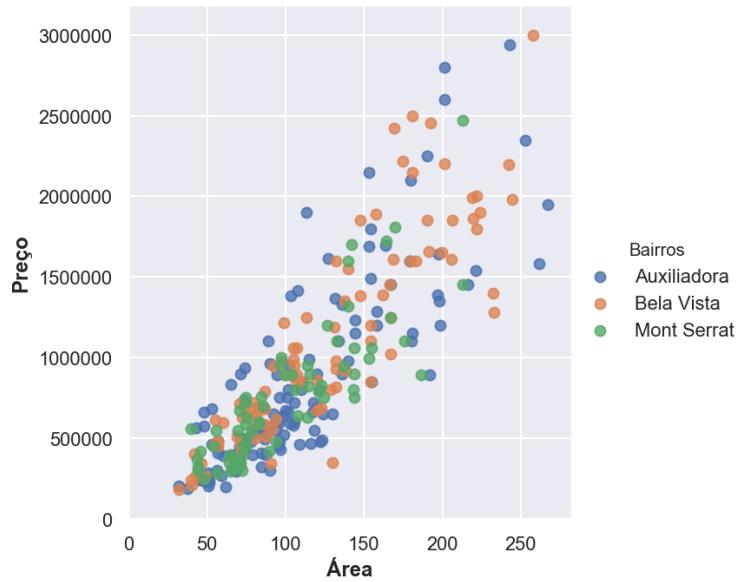
Utilizando da análise multivariada, o gráfico a seguir mostra as relações dos dados entre as variáveis com destaque para os bairros. Nele é possível visualizar que os dados do bairro Bela Vista (cor rosa) se destacam dos outros dois, sempre aparecendo mais à direita. Isso é causado pelos seus imóveis possuírem preços maiores, áreas maiores, valores de condomínio maiores e valores de IPTU maiores também.

Gráfico 15 – Pairplot: relacionamento entre todas as variáveis destacando os bairros



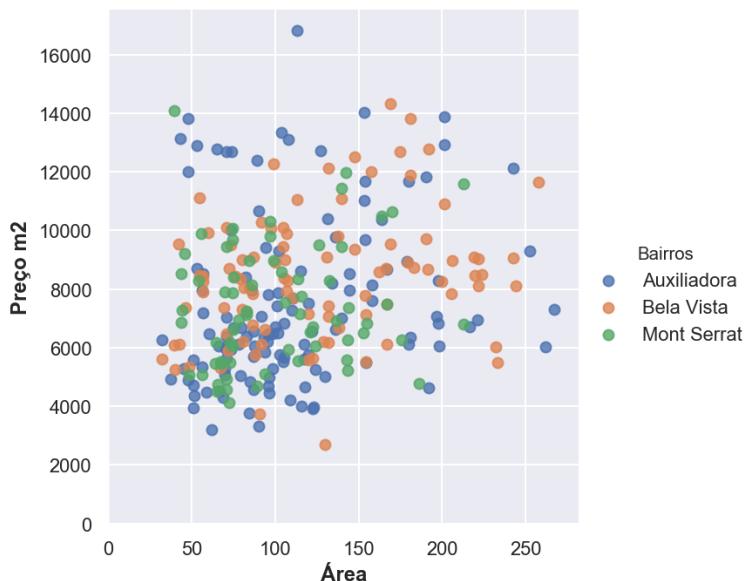
No gráfico seguinte, busca-se dar destaque para a relação área, preço e bairro. Numa visualização maior que o *pairplot* anterior, observa-se que os pontos laranjas, de Bela Vista, se destacam ocupando o quadrante direito superior.

Gráfico 16 – Distribuição dos imóveis por área x preço x bairro



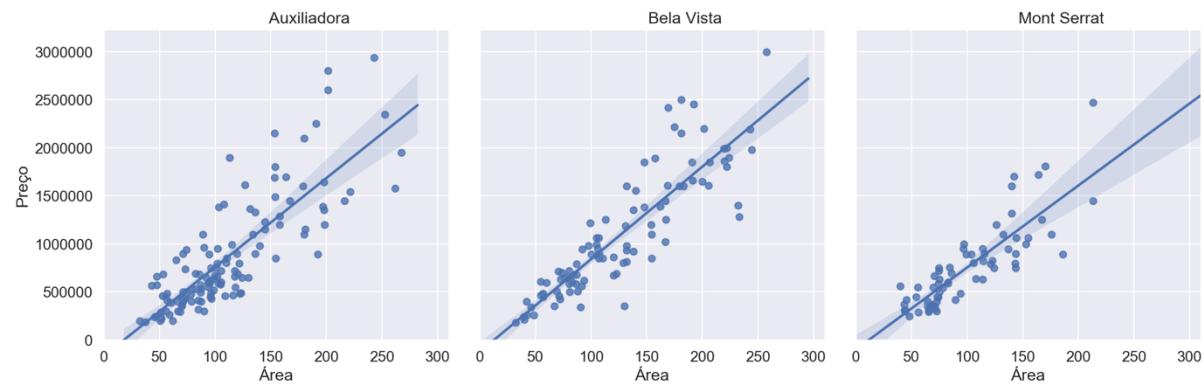
Analisando graficamente (Gráfico 17) a mesma distribuição de pontos, mas utilizando o preço do m^2 no eixo y, ao invés do preço do apartamento, uma informação importante surge. É possível observar que há apartamentos do mesmo tamanho e no mesmo bairro com preços do m^2 diferentes. Uma hipótese para essa questão seria que apartamentos novos geralmente têm preços do m^2 maior que os usados.

Gráfico 17 – Distribuição dos imóveis por área x preço do m^2 x bairro



E no Gráfico 18 tem-se um conjunto dos três *plots* comparando a relação área e preço separadamente para cada bairro.

Gráfico 18 – Regressão linear entre as várias área e preço com agrupamento por bairro



4. Resultados

Após o tratamento inicial dos dados e análise exploratória, o trabalho atuou, no objetivo principal, em desenvolver um modelo preditivo para o preço de um apartamento. De posse de sete atributos e mais a variável “preço” (*target* do modelo), determinou-se, primeiramente, um modelo de *benchmark* com o uso do algoritmo de Regressão Linear. Os atributos utilizados foram:

1. Área do apartamento ('area')
2. Valor do condomínio ('condominium')
3. Valor do IPTU ('iptu')
4. Quantidade de vagas de garagem ('box')
5. Quantidade de quartos ('rooms')
6. Bairro Bela Vista ('district_Bela Vista')
7. Bairro Mont'Serrat ('district_Mont Serrat')

O modelo de *Benchmark* de Regressão Linear teve um bom resultado com o R^2 , cujo as estatísticas de medição do seu desempenho, com o uso do conjunto de dados de teste, foram:

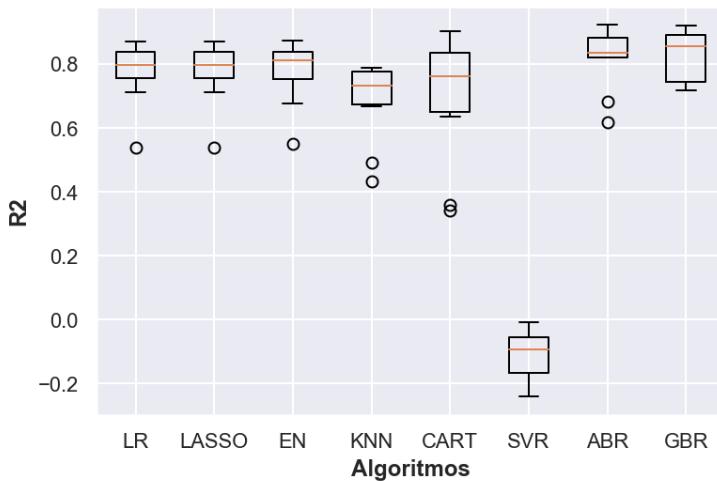
- MSE do benchmark: 74,061,685,914.85
- MAE do benchmark: 200,770.87
- R^2 do benchmark: 0.75

Em seguida, oito algoritmos regressores foram executados nas suas configurações padrões utilizando os dados de treinamento. Com exceção do SVR, todos os demais apresentaram resultados próximos entre si. O Gradient Boosting Regressor obteve a melhor pontuação, com R^2 igual a 0,83. As pontuações foram:

LR : Linear Regression =	0,77
LASSO : Lasso =	0,77
EM : Elastic Net =	0,78
KNN : KNeighbors Regressor =	0,68
CART : Decision Tree Regressor =	0,69
SVR : Support Vector Regression =	-0,11
AB : AdaBoost Regressor =	0,82
GBR : Gradient Boosting Regressor =	0,83

O gráfico a seguir compara o desempenho de cada algoritmo.

Gráfico 19 – Comparação dos Algoritmos



O passo seguinte foi refinar os parâmetros do GBR no intuito de se obter um modelo ainda superior. Partiu-se para o uso do GridSearchCV, cujos parâmentros iniciais foram:

```
param_grid = {'max_depth':range(5,16,2),
              'max_features':range(1,len(features_list_dum),1),
              'min_samples_leaf':range(30,71,10),
              'min_samples_split':range(200,1001,200),
              'n_estimators':range(20,81,10)
             }
```

Depois de ajustes, a configuração final para a busca do GridSearchCV foi:

```
param_grid = {'max_depth':range(1,10,2),
              'max_features':range(1,len(features_list_dum),1),
              'min_samples_leaf':range(2,10,2),
              'min_samples_split':range(2,20,4),
              'n_estimators':range(2,81,10)
             }
```

Com essa configuração de busca o GridSearchCV encontrou a parametrização otimizada para o GBR, no conjunto de treinamento, resultando no R2 igual a 0,86. Os parâmetros otimizados foram:

```
{'max_depth': 9, 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 32}
```

Com esse resultado, seguiu-se para a configuração do modelo final usando o GBR. Primeiro gerou-se o modelo preditivo com o conjunto de dados de treinamento. Em seguida, executou-se o modelo criado do GBR com os dados de teste. As estatísticas de avaliação do modelo foram:

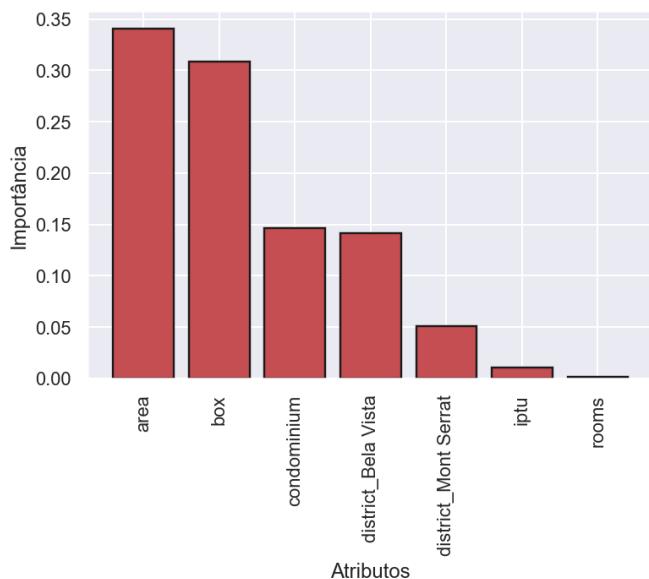
- MSE: 54,877,748,414.29
- MAE do benchmark: 166,155.67
- R2: 0.82

O R^2 , de 0,82, obtido com o conjunto de teste, ficou próximo ao de 0,86 alcançado com os dados de treinamento. Com isso, entende-se que não houve *overfitting*. Na comparação do GBR com o modelo de *benchmark*, o GBR foi superior nas três estatísticas de avaliação, MSE, MAE e R^2 .

Foram identificados os graus de importância de cada atributo utilizado no modelo GBR. A área do apartamento e quantidade de vagas de garagem foram os dois principais atributos, somando 65% da importância dos atributos. O valor do condomínio vem em seguida com 15%. E IPTU e a quantidade de quartos foram os dois últimos, com 1% e 0%, respectivamente. O gráfico a seguir apresenta os atributos e suas importâncias na forma de barra. Os valores das contribuições de cada atributo para o modelo foram:

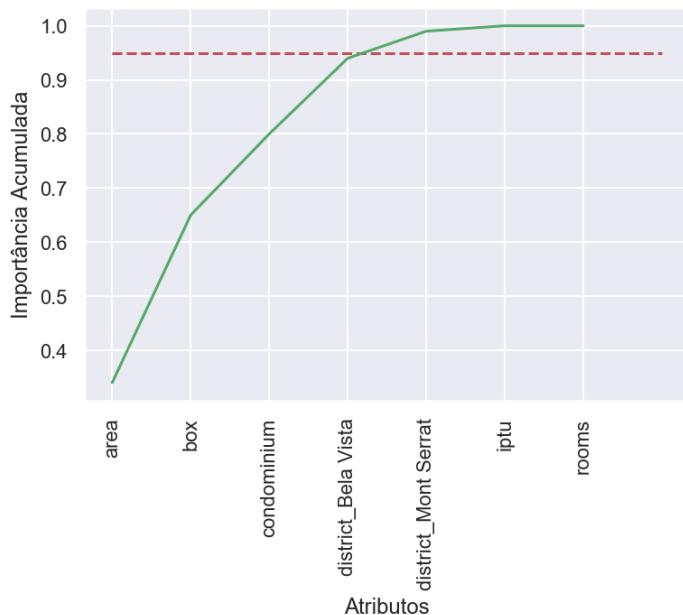
Atributo: area	Importância: 0.34
Atributo: box	Importância: 0.31
Atributo: condominium	Importância: 0.15
Atributo: district_Bela Vista	Importância: 0.14
Atributo: district_Mont Serrat	Importância: 0.05
Atributo: iptu	Importância: 0.01
Atributo: rooms	Importância: 0.0

Gráfico 20 – Importância dos Atributos



O Gráfico 21, por sua vez, apresenta a importância acumulada dos atributos do modelo. Observa-se que linha tracejada em vermelho indica o nível de 95%, de forma que possa ser visualizado quando a curva cruza esse nível.

Gráfico 21 – Importância Acumulada



Um objetivo específico elencado no início do trabalho foi buscar prever o preço de um apartamento. Desta forma, tendo concluído o modelo final, estimou-se que um apartamento, de 119m², com 2 vagas de garagem, com R\$750,00 de valor de condomínio, com R\$2.100,00 de valor de IPTU, com 3 quartos e localizado no bairro de Mont'Serrat, tem o seu valor previsto R\$713.420,00.

5. Conclusão

5.1. Resultados obtidos

Este trabalho se propôs a prever o preço de venda de um apartamento na cidade de Porto Alegre. Foram elencados ainda dois objetivos secundários: identificar o melhor algoritmo para previsão e identificar os atributos mais importantes. A primeira etapa do trabalho consistiu na coleta de dados no site da imobiliária Foxter por meio de um *web scaper* construído para essa finalidade. Na segunda etapa os dados foram lidos do csv, limpos e formatados. A análise exploratória foi feira na terceira etapa. E, por fim, a quarta etapa foi a modelagem.

Foram identificados e utilizados oito algoritmos regressores. Um, o de regressão linear, foi selecionado para ser o *benchmark* do trabalho. Ao final dos testes, o Gradient Boosting Regressor obteve o melhor desempenho dentre todos, e teve um R^2 com 0,82 com o modelo final. Seu resultado foi melhor inclusive que o do *benchmark*.

Os atributos relativos ao tamanho dos apartamentos e a quantidade de vagas de garagem foram identificados como os dois principais atributos para o modelo. Suas importâncias para o modelo final foram de 34% e 31%, respectivamente, cuja soma totaliza em 65%. Por sua vez, o valor do IPTU e a quantidade de quartos foram os de menor importância.

5.2. Aperfeiçoamento do modelo

Dois aspectos foram identificados durante o trabalho que podem ser explorados futuramente na busca pelo aperfeiçoamento do modelo. Durante a análise exploratória, verificou-se que há apartamentos com o mesmo tamanho, no mesmo bairro, mas com valores distintos de preço do m². A identificação de possíveis causas e sua categorização permitirá separar os imóveis e, com isso, se ter uma melhor explicação para diferença de preços. O segundo aspecto seria o tratamento do texto incluso no campo “descrição do imóvel” no site da imobiliária. O conjunto das palavras (normalmente positivas sobre o imóvel) podem vir a contribuir para uma melhor explicação da formação do preço.

6. Referências Bibliográficas

ALFIYATIN, Adyan Nur, FEBRITA, Ruth Ema, TAUFIQ, Hilman, MAHMUDY, Wayan Firdaus. (2017). **Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study**: Malang, East Java, Indonesia International Journal of Advanced Computer Science and Applications (IJACSA), 8(10). <https://bit.ly/2FzdiHf>

BALDOMINOS, Alejandro, BLANCO, Iván, MORENO, Antonio José, ITURRARTE, Rubén, BERNÁRDEZ, Óscar, AFONSO, Carlos. (2018). **Identifying Real Estate Opportunities Using Machine Learning**. *Appl. Sci.* 2018, 8, 2321. <https://bit.ly/2QVRTTA>

BHATTACHARYYA, Indresh. (2018). **Support Vector Regression Or SVR**. <https://bit.ly/2Sk04Ow>

BROWNLEE, Jason. (2014). **Why you should be Spot-Checking Algorithms on your Machine Learning Problems**. <https://bit.ly/2CNwruW>

BROWNLEE, Jason. (2016). **Boosting and AdaBoost for Machine Learning**. <https://bit.ly/2MFBQsG>

BROWNLEE, Jason. (2018). **Machine Learning Mastery with Python**: understand your data, create accurate models and work projects end-to-end. E-Book. <https://bit.ly/2MDN6WA>

KANDAN, Harish. (2017). **Understanding the kernel trick**. Towards Data Science. 30/08/2017. <https://bit.ly/2B9MvqS>

KOMAGOME-TOWNE, Anh. (2016). **Models and visualizations for housing price prediction**. California State Polytechnic University. Thesis. <https://bit.ly/2swsg1S>

MÜLLER, Andrea C. & GUIDO, Sarah. (2017). **Introduction to Machine Learning with Python**: a guide for data scientists. O'Reilly.

NGUYEN, An. (2018). **Housing Price Prediction**. Union College. Final Project. <https://bit.ly/2RO6lSf>

PARK, Byeonghwa, BAE, Jae Kwon. (2015). **Using machine learning algorithms for housing price prediction**: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, Volume 42, Issue 19, 1 November 2015, Pages 6806. <https://bit.ly/2Cqxlxo>

PIERRE, Rafael. (2018). **Going Dutch: How I Used Data Science and Machine Learning to Find an Apartment in Amsterdam**—Part I. Towards Data Science. <https://bit.ly/2HI3ReZ>

SWALIN, Alvira. (2018). **Choosing the Right Metric for Evaluating Machine Learning Models(Part 1)**. <https://bit.ly/2HwlUtT>

YU, Li, JIAO, Chenlu, XIN, Hongrun, WANG, Yan, WANG, Kaiyang. (2018). **Prediction on Housing Price Based on Deep Learning**. International Journal of Computer and Information Engineering Vol:12, No:2, 2018. <https://bit.ly/2MhHNfg>