Modelo para estimativa de preço de lançamento para apartamentos no estado de São Paulo

ANALYTICS E INTELIGÊNCIA ARTIFICIAL

Nome do Aluno:
Laerton Amorim Correia

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    - 3.1. Bases originais
    - 3.2. Filtros
    - 3.3. Principais Variáveis
    - 3.4. Descrição das Variáveis
    - 3.4. Processo de redução de variáveis
4. Análise Exploratória dos Dados
5. Modelagem com estatística
6. Modelagem com inteligência Artificial
7. Conclusões
8. Sugestão para Trabalhos Futuros

1. Objetivo do Trabalho

Este trabalho tem o objetivo estimar o preço de venda de apartamentos residenciais novos. Este valor é sugerido como preço de lançamento para cada unidade habitacional do empreendimento, tal valor é obtido considerando características físicas e geográficas de cada unidade habitacional.

Observações:

A expectativa é que o modelo final obtido neste trabalho possa auxiliar possíveis compradores e/ou investidores para a aquisição/investimento em imóveis no estado de são Paulo.

A base de dados utilizada foi obtida no site: https://centrodametropole.fflch.usp.br/pt-br/download-de-dados?download_dados=embraesp&items_per_page=20

2. Contextualização do Problema

No mercado imobiliário cada consumidor monta sua cesta de necessidades básicas a partir de suas preferências em relação às características presentes em cada bem. Dessa forma, na análise do preço de um imóvel sãosomente as características físicas do imovel são significantes mas tambem efeitos de externalidades devem ser considerados.

A estrutura espacial do ambiente construído contribui para a valorização dos imóveis. As características gerais e de funcionalidades dessa estrutura são denominadas de externalidades.

Os modelos de precificação de imóveis tentam explicar o comportamento de um tipo de mercado onde se transacionam bens com atributos diferentes.

A quantidade de atributos que o bem possui reflete no preço que equilibra esse mercado. Quanto melhores os atributos, maior o preço a eles atribuído.

A avaliação de imóveis e de seus preços pode ser realizada de forma mais assertiva com o uso de modelos de regressão.

3. Bases de Dados

Base 1:

Base de Lançamentos Imobiliários Residenciais na Região Metropolitana de São Paulo (1985-2013)

Base 2:

Base Cartográfica Digital Georreferenciada das Linhas e Estações de Trem e Metrô - Transporte sobre Trilhos - Região Metropolitana de São Paulo 2021

3.1. Bases Originais

Base 1:

Histórico dos dados

A base de dados apresenta Dados relativos a aperfeiçoamento qualitativo e ampliação de informações sobre os lançamentos imobiliários residenciais da Região Metropolitana de São Paulo, sistematizadas pela Empresa Brasileira de Patrimônio (Embraesp) no período entre 1985 e 2013.

Processo de captação dos dados

O CEM (com apoio financeiro Cepid/Fapesp e Inct/CNPq) realizou o aperfeiçoamento qualitativo e ampliou as informações a respeito dos lançamentos imobiliários residenciais da RMSP, sistematizados pela Empresa Brasileira de Patrimônio(Embraesp). Estes dados foram disponibilizados pelo Centro de estudos da Metrópole.

Visão da Base

A base exibe informações importantes para caracterização do preço final do imóvel.

3.1. Bases Originais

Visão da base: a base de dados total é composta por quatro arquivos e um dicionário de dados, foram utilizados os arquivos .DBF que contém uma tabela com 16935 linhas e 85 colunas e .shp com uma coluna e mesma quantidade de linhas.

O primeiro descreve lançamentos de empreendimentos residenciais no período citado e o segundo contém a projeção Cartográfica dos dados.

LanRes_85_13_RMSP_CEM.prj
LanRes_85_13_RMSP_CEM.shp
LanRes_85_13_RMSP_CEM.shx
LANRES_85_13_RMSP_CEM.DBF
Dicionário_Lan_Res_1985-2013_RMSP_CEM.pdf

3.1. Bases Originais

Base 2:
Histórico dos dados

Este conjunto de arquivos resume a infraestrutura de transporte de passageiros sobre trilhos na RMSP. Concluídos em março de 2021, foram elaborados com base no arquivo CEM de Logradouros, com o apoio de imagens de satélite.

Processo de captação dos dados

As informações foram coletadas principalmente nos sítios oficiais da Companhia do Metropolitano de São Paulo - Metrô, da Companhia Paulista de Trens Metropolitanos e no sítio www.estacoesferroviarias.com.br. Também foram consultados outros sítios, como o da Secretaria Estadual dos Transportes Metropolitanos e o Google Maps, além dos guias impressos Geomapas e Mapograf. Na denominação dos arquivos, "L" refere-se a linha, "E" a estação (ou terminal). Dois dos arquivos, porém, extrapolam os limites da RMSP - exceção adotada para incluir completamente o sistema de trens metropolitanos da CPTM, que chegam a Jundiaí.

Visão da Base

A base exibe informações de Infraestrutura de transporte de passageiros sobre trilhos - RMS

3.1. Bases Originais

Visão da base: a base de dados total é composta por quatro arquivos e um dicionário de dados, foi utilizado o arquivo MetE_2021_CEM que contém 198 linhas e 12 colunas.

MetL_2021_CEM - Linhas de metrô e trem na Região Metropolitana de São Paulo

MetE_2021_CEM - Estações de metrô e trem na Região Metropolitana de São Paulo

MetLex_2021_CEM - Expansão em andamento das linhas de metrô e trem na Região Metropolitana de São Paulo

MetEex_2021_CEM - Expansão em andamento das estações de metrô e trem na Região Metropolitana de São Paulo

3.2. Filtros

Base Original
16935 imóveis e 85 variáveis

Base de Partida

A Base original arquivo .DBF com 16935 imóveis e 85 variáveis.

Base Filtrada
14599 imóveis e 85 variáveis

Redução da Base

Foram removidos imóveis(16) com informações com falta de informação.

Base Analítica
14599 imóveis e 23 variáveis

Redução das Variáveis

As informações do arquivo .shp foram inseridas no arquivo .dbf. Os dados do arquivo MetE_2021_CEM foram utilizados para cálculos. Foram selecionadas as variáveis pertinentes

Base Final
11679 imóveis 23 variáveis

Base final

A base foi dividida em 11679 imóveis para a modelagem e 2920 imóveis para verificar o modelo(foram selecionados somente imóveis verticais)

3.3. Principais variáveis

Variáveis numéricas

- garagens
- area_util
- area_total
- dormitorios
- banheiros
- unidades_por_andar
- andares
- distancia_trem_metro
- preco_m2_atualizado

Variáveis categóricas

- sistema_finaceiro
- setor
- zona
- municipio
- cooperativa
- hotel
- flat
- exflat
- estacao

Variáveis de chave

- data_ent
- data_lan
- ano_lan

Variável target

- preco_de_venda_atualizado2013

3.4. Descrição das variáveis

Variáveis numéricas

garagens: quantidade de vagas de garagem para cada apartamento
area_util: área útil por apartamento em metros quadrados
area_total: área total por apartamento em metros quadrados
dormitorios: quantidade de dormitórios para cada apartamento
banheiros: quantidade de banheiros para cada apartamento
unidades_por_andar: quantidade de apartamentos para cada andar do prédio
andares: quantidade de andares para cada prédio
distancia_trem_metro: distancia em quilometros do apartamento até a estação de trem ou metrô mais próxima(obtido via cálculo)
preco_m2_atualizado: preço por metro quadrado da área útil do apartamento na época lo lançamento atualizado em valores de dezembro de 2013 pelo IGP-DI

Variáveis categóricas

sistema_finaceiro: sistema financeiro adotado(preço de custo/preço fechado)
setor: setor do empreendimento no mapa oficial da cidade
zona: zona do empreendimento pelo mapa oficial da cidade como as ZMp(zona mista de proteção ambiental em zonRev2.qxd (prefeitura.sp.gov.br))
municipio: cidade onde se encontra o apartamento
cooperativa: o apartamento foi construído por cooperativa (1-sim/0-não)
hotel: caracteriza se o emprendimento é um hotel(1-sim/0-não)
Flat: caracteriza se o emprendimento é um flat(1-sim/0-não)
Exflat: Foi contruído para ser flat, mas agora é um residencial comum(1-sim/0-não)
Estacao: estação de trem ou metrô mais próxima do apartamento

3.4. Descrição das variáveis

Variáveis chave

data_ent: data em que o empreendimento foi entregue ou previsão de entrega
data_lan: data em que o empreendimento foimlançado
ano_lan: ano em que o empreendimento foi lançado

Variável Target

preco_de_venda_atualizado2013: preço de venda da unidade na época do lançamento atualizado em valores de dezembro de 2013 pelo IGP-DI

4. Análise exploratória de dados

Metodologia:

- A base de dados contempla o período temporal de janeiro de 1985 a setembro de 2013.
- Como a base possui a princípio 2.326 imóveis horizontais e 14.609 imóveis verticais foi decidido fazer a análise somente com verticais.
- Considerando que a base de dados contempla lançamentos em por 29 anos teríamos em média aritimética aproximadamente 503 observações ao ano, sendo esta quantidade de registros insuficientes para uma análise aceitável.
- Por este motivo a variável target escolhida será o preço de venda da unidade no seu lançamento atualizado em valores de dezembro de 2013 pelo IGP-DI .
- Assim pode-se usar o total de registros de verticais para precificação no ano de 2013.

4. Análise exploratória de dados

Variáveis Qualitativas: análise univariada

- sistema_financeiro , setor, zona, municipio, cooperativa, hotel, flat, exflat, tipo_de_empreendimento, estacao

A tabela 1 exibe informações de frequência para as variáveis categóricas.

Observando a tabela 1 verificamos que o perfil predominante dos registros da base de dados são imóveis adquiridos por sistema de preço fechado, localizados na cidade de São Paulo, construídos e utilizados como imóveis residenciais e não construídos por cooperativa.

Tabela1: frequência das variáveis categóricas

	sistema financeiro	setor	zona	município	cooperativa	hotel	flat	exflat	estação
categorias	3	205	54	25	2	2	2	2	135
top categoria	preço fechado	m	2	São Paulo	não	não	não	não	Pref. Celso Daniel
top frequência	13277	3289	3759	11374	14400	14535	14351	14594	735

4. Análise exploratória de dados

Variáveis Quantitativas: análise univariada

- garagens, area_util, area_total , dormitorios, banheiros, unidades_por_andar, andares, distancia_trem_metro, preco_m2_atualizado

A tabela 2 exibe as medidas de dispersão e posição para a variáveis qualitativas.

Observando a tabela 2 verifica-se 75% dos imóveis custam até R$731.023,15, estes imóveis estão relativamente próximos a estações de metrô ou trem até 2,75 km.

Pode-se verificar que o preço de venda assim com as demais variáveis apresentam grande amplitude e dispersão.

Tabela2: Medidas de posição e dispersão

	mean	std	min	25%	50%	75%	max
garagens	2	1.08	0	1	2	2	12
area_util(m2)	100	77.75	13	56	72	118	1975
area_total(m2)	187	149.51	30	101	134	220	4000
dormitorios	3	0.90	1	2	3	3	6
banheiros	2	0.89	1	1	2	2	6
unidades_por_andar	5	3.04	0	3	4	6	60
andares	15	6.24	0	10	14	19	46
distancia_trem_metro(km)	2	1.50	0.0	0.8	1.5	2.8	12.3
preco_m2_atualizado(R$)	5946.01	2979	944	3949	5285	7097	43431
preco_de_venda_atualizado2013(R$)	671161.00	958361	65930	250693	402863	731023	31843098

4. Análise exploratória de dados

Variáveis Quantitativas: análise univariada

A tabela 3 exibe a variação dos dados obtidos em relação à média para cada variável quantitativa. Esta medida é conhecida como **coeficiente de variação.**

Observando a tabela 3 é verifcado que os coeficientes de variação apresentam valor alto o que sugere grande variação dos dados.

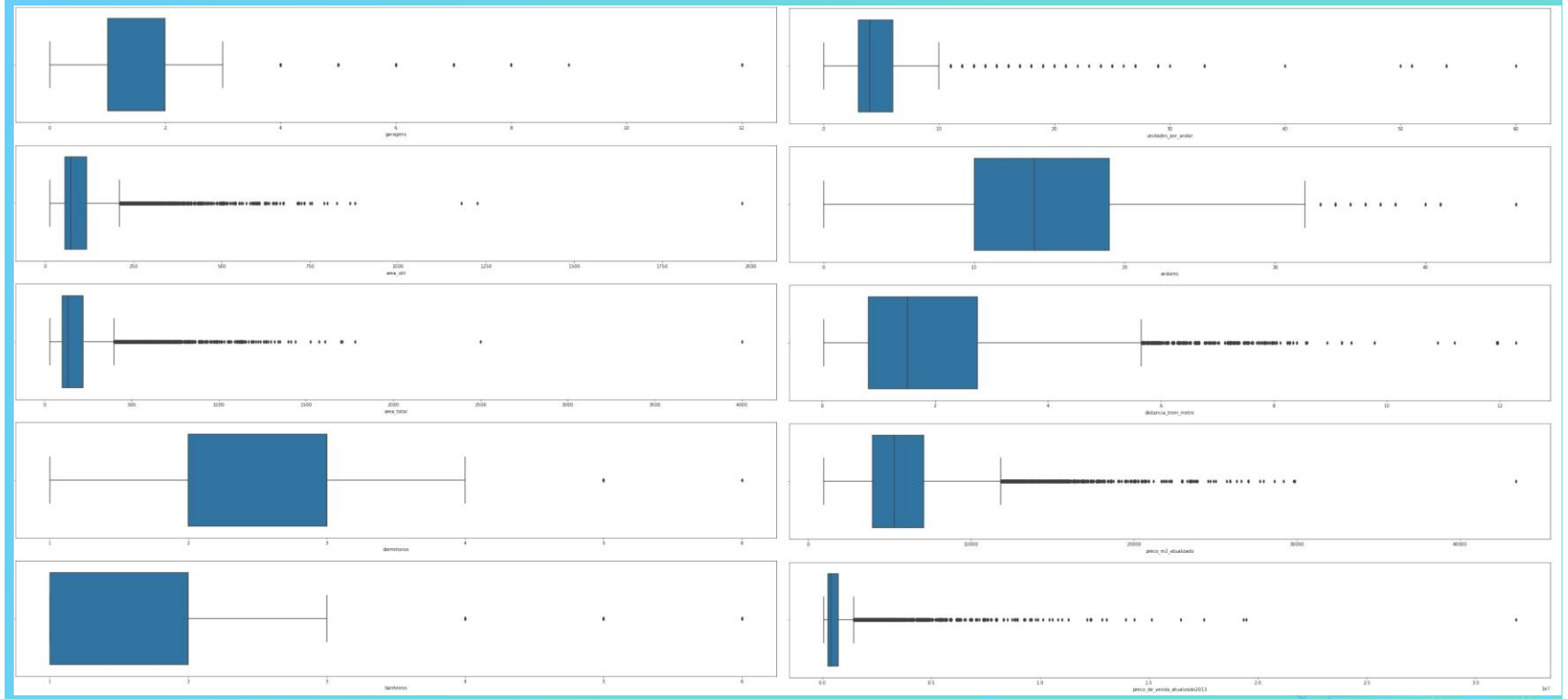
Tabela3: coeficientes de variação

Variáveis	Coeficiente de Variação
garagens	0.60
area_util	0.78
area_total	0.80
dormitorios	0.34
banheiros	0.46
unidades_por_andar	0.66
andares	0.43
distancia_trem_metro	0.77
preco_m2_atualizado	0.50
preco_de_venda_atualizado2013	1.43

4. Análise exploratória de dados

Variáveis Quantitativas: análise univariada

Observando a figura 1, verifica-se que as variáveis apresentam grande quantidade de outliers, contudo isso pode ser melhorado.



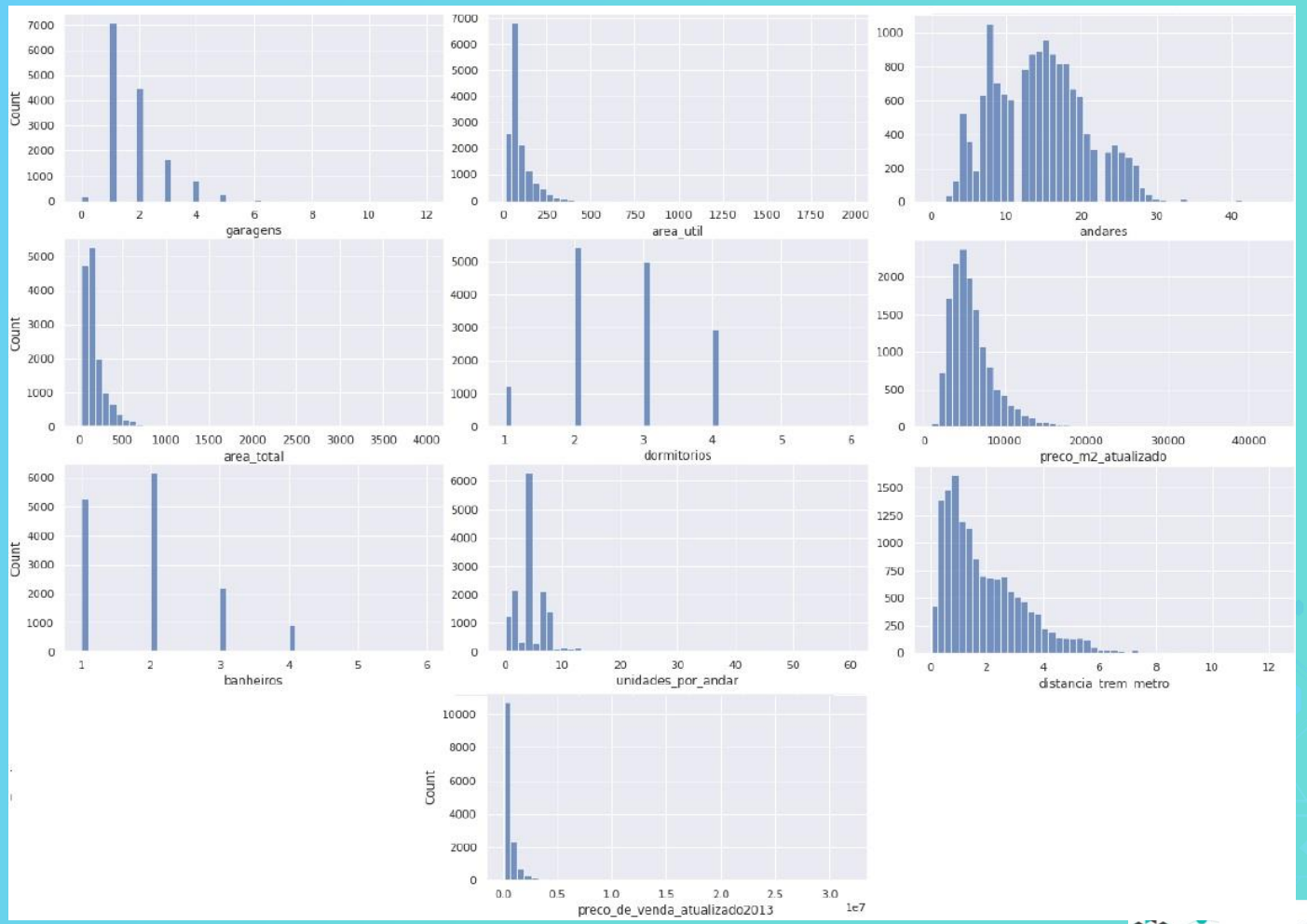
4. Análise exploratória de dados

Com tentar diminuir a quantidade de outtliers???

4. Análise exploratória de dados

Variáveis Quantitativas: análise univariada

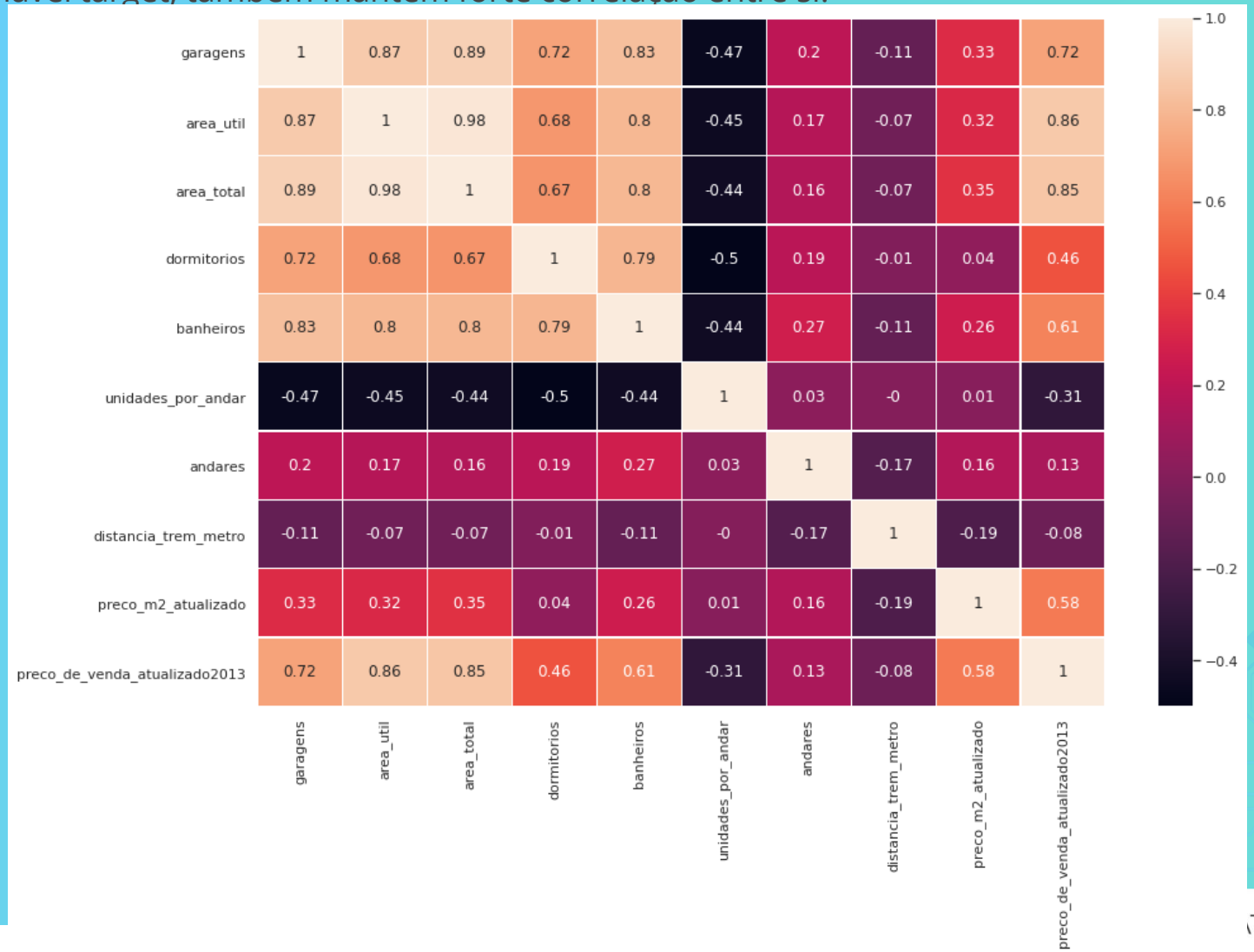
Observando a figura 2 verifica-se, com exceção da variável andares, que as variáveis não seguem uma distribuição nomal sendo em geral assimétricas à direita.



4. Análise exploratória de dados

Variáveis Quantitativas x Variável Target: análise bivariada

Observando a correlação das features entre sí e com a variável target pode-se observar que embora algumas variáveis como area_util e area_total mantenham correlação forte com a variável target, tambem mantém forte correlação entre sí.



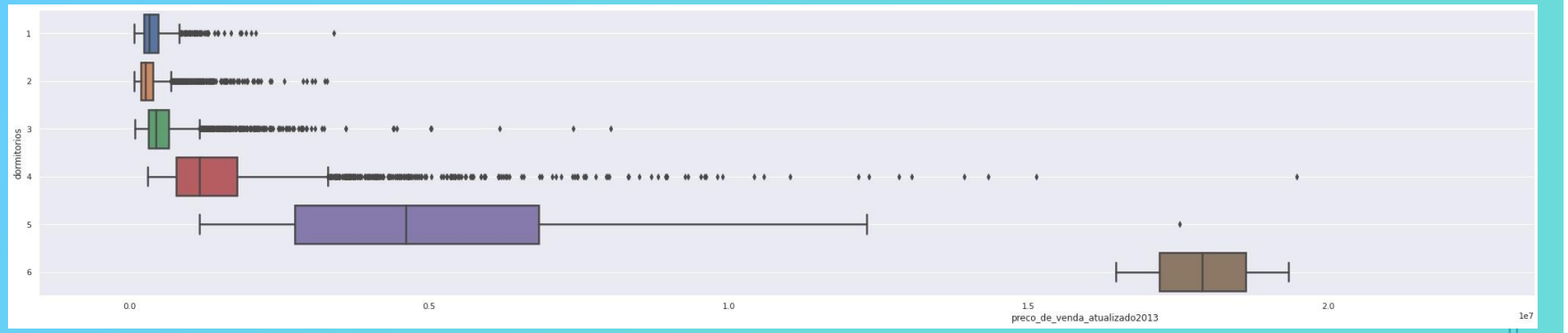
4. Análise exploratória de dados

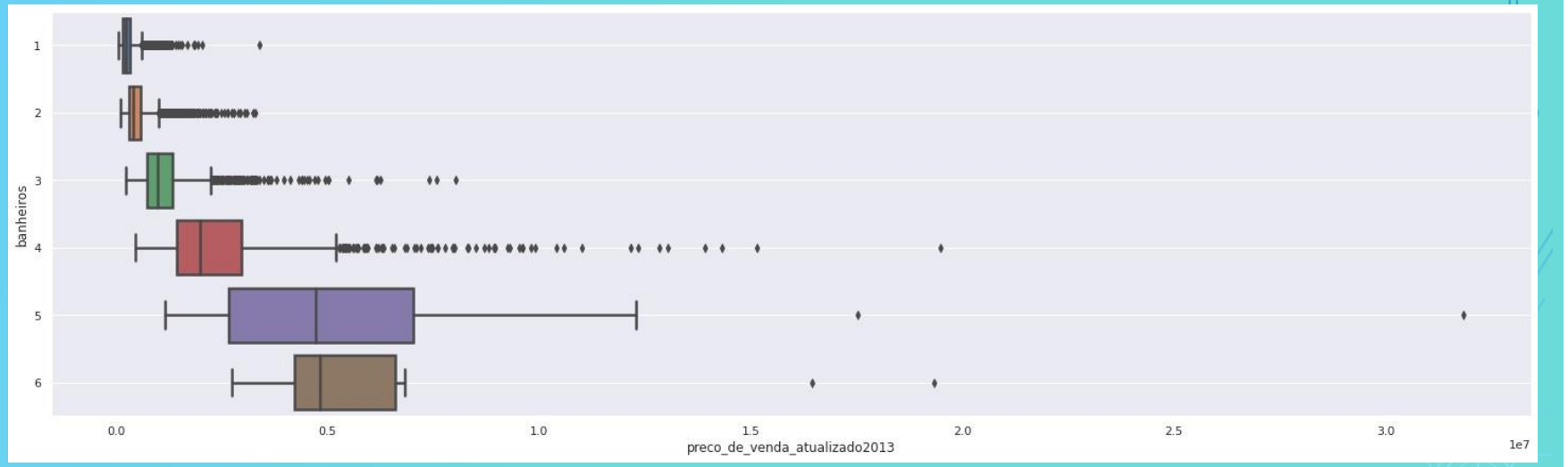
Como eliminar ou diminuir a correlação das features entre sí e manter a correlação das features com a variável target?

4. Análise exploratória de dados

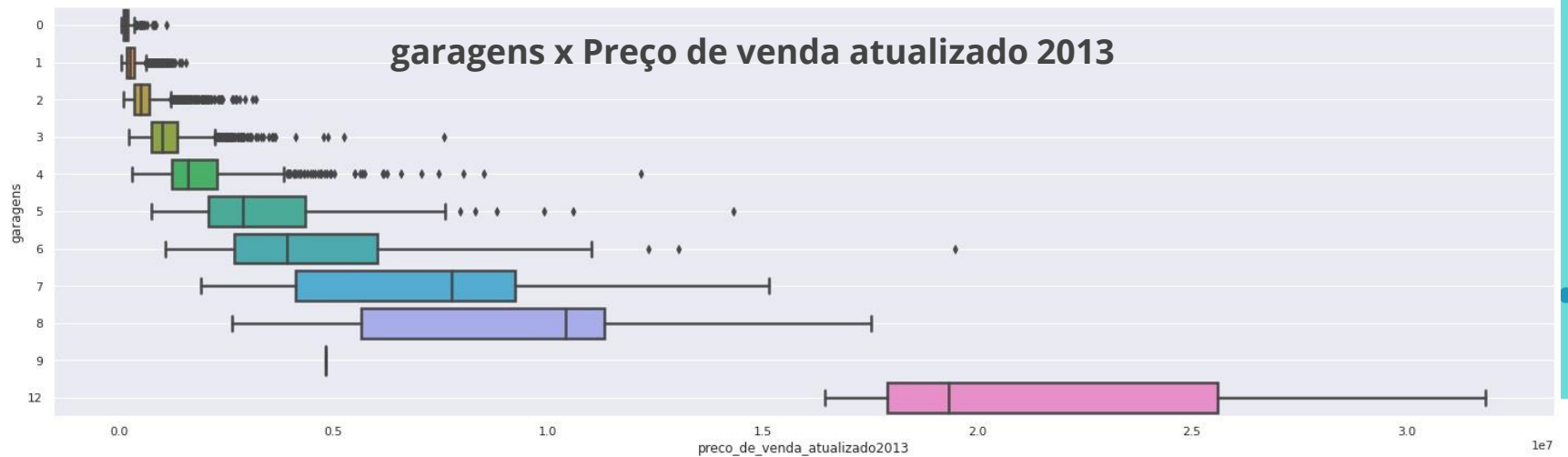
Variáveis Quantitativas x Variável Target: análise bivariada

Nos grafícos podemos observar a tendência de aumento da mediana do preço de venda atualizado com para apartamentos com mais dormitórios, banheiros e garagens.





4. Análise exploratória de dados



Variáveis Quantitativas
X
Variável Target:

análise bivariada

Nos grafícos podemos observar a tendência de aumento da mediana do preço de venda atualizado com para apartamentos com mais dormitórios, banheiros e garagens.

5. Modelagem com estatística

O objetivo neste primeiro momento é fazer a precificação com um modelo de regressão linear multipla.

Existem duas questões importantes até aqui:

- Como minimizar a quantidade de outliers?

- Como minimizar a correlação das features entre sí e manter a correlação das features com a variável target?

5. Modelagem com estatística

Como diminuir a quantidade de outliers ?

Uma possível saída é a segmentação da base de dados.

Contudo como efetuar esta segmentação?

Foram efetuadas tentativas de segmentação por métrodo dos quartis e por clusterização via kmeans e pelo métrodo Elbow. Estas abordagens não foram eficientes para solução do problema.

Contudo olhando com mais atenção para a variável zona...

5. Modelagem com estatística

Segmentação da base de dados

Para segmentar vamos olhar para a variável zona.

O que é Zoneamento Urbano?

O zoneamento urbano é um plano que divide um determinado espaço (cidade) em zonas territoriais e determina, para cada uma delas, as regulamentações pertinentes quanto ao uso e ocupação do solo.

Essa divisão da cidade em zonas leva em consideração diferentes fatores.

Como exemplo pode-se citar a função predominante pretendida para cada região, comercial, industrial, residencial, mista ou até a pretensão de preservar patrimônios naturais ou áreas de interesse cultural e a manutenção de centros históricos

Macrozona de Estruturação e Qualificação Urbana	
ZER - 1	zona exclusivamente residencial de densidade demográfica baixa
ZER - 2	zona exclusivamente residencial de densidade demográfica média
ZER - 3	zona exclusivamente residencial de densidade demográfica alta
ZPI	zona predominantemente industrial
ZM - 1	zona mista de densidades demográfica e construtiva baixas
ZM - 2	zona mista de densidades demográfica e construtiva médias
ZM - 3a	zona mista de densidades demográfica e construtiva altas
ZM - 3b	zona mista de densidades demográfica e construtiva altas
ZCPa	zona centralidade polar de densidades demográficas e construtiva médias
ZCPb	zona centralidade polar de densidades demográficas e construtiva altas
ZCLa	zona centralidade linear de densidades demográficas e construtiva médias
ZCLb	zona centralidade linear de densidades demográficas e construtiva médias
ZCLz - I	zona centralidade linear destinada à localização das atividades de comércio e serviços de baixa densidade
ZCLz - II	zona centralidade linear destinada à localização das atividades de serviços de baixa densidade
ZOE	zona de ocupação especial

5. Modelagem com estatística

Segmentação da base de dados

Em uma análise via boxplot do comportamento das categorias da variável zona.

Verifica-se a relação entre preco_venda_atualizado2013 e zona, agupando a mediana do preço por zona e ordenando de forma crescente fornece “um ranqueamento” das zonas de acordo com as citadas medianas.

O conjunto de zonas foi dividido em três grandes grupos e chamados de zona1, zona2 e zona3.

Zonas de cada segmento:

- **Zona 1**:

zcl-a, 07, zo, 09, p, zeis, zmp, 11, zm, zcpb2, 08

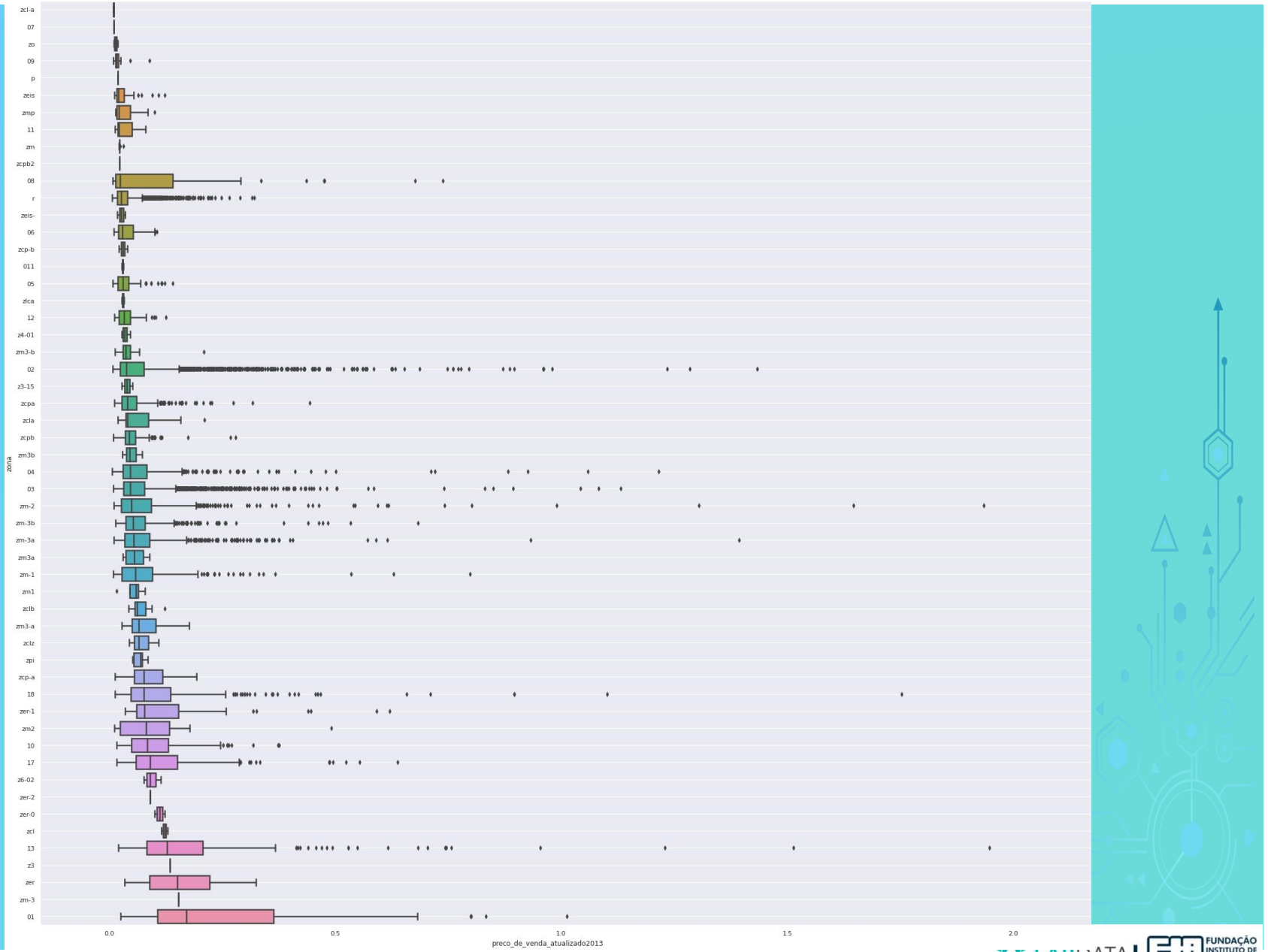
- **Zona 2**:

r, 06, zcp-b, 011, 05, zlca, 12, z4-01, zm3-b, 02, z3-15, zcpa, zcla, zcpb, zm3b, 04, 03, zm-2, zm3a, zm-3b, zm-3a, zm-1, zm1, zclb, zm3-a, zclz

- **Zona 3:**

zpi, zcp-a, 18, zer-1, zm2, 10, 17, z6-02, zer-2, zer-0, zcl, 13, z3, zer, zm-3, 01

5. Modelagem com estatística



5. Modelagem com estatística

Como eliminar ou diminuir a correlação das features entre sí e manter a correlação das features com a variável target?

A primeita tentativa foi a de criar novas features combinando as variáveis numéricas que possímos afim de que estas nova features tivessem menor dependência lienar entre sí. Contudo esta abordagem não foi bem sucedida.

Em uma nova abordagem optou-se por trabalhar com transformaçãos de variáveis afim de tentar “linearizar o modelo”.

Olhando para a segmentação zona1

5. Modelagem com estatística

Transformação de variáveis

Para a variável target optou-se porm uma transformação logarítimica por sua melhor adequação a distribuição da variável e por simplicidade de obtenção de sua inversa.
Observação: no python log1p(x)

Para cada uma das demais variáveis numéricas foram utilizadas as seguintes transformaçãoes:

Gauss
Clipping
Box Cox
Yeo Johnson
Logarítmica

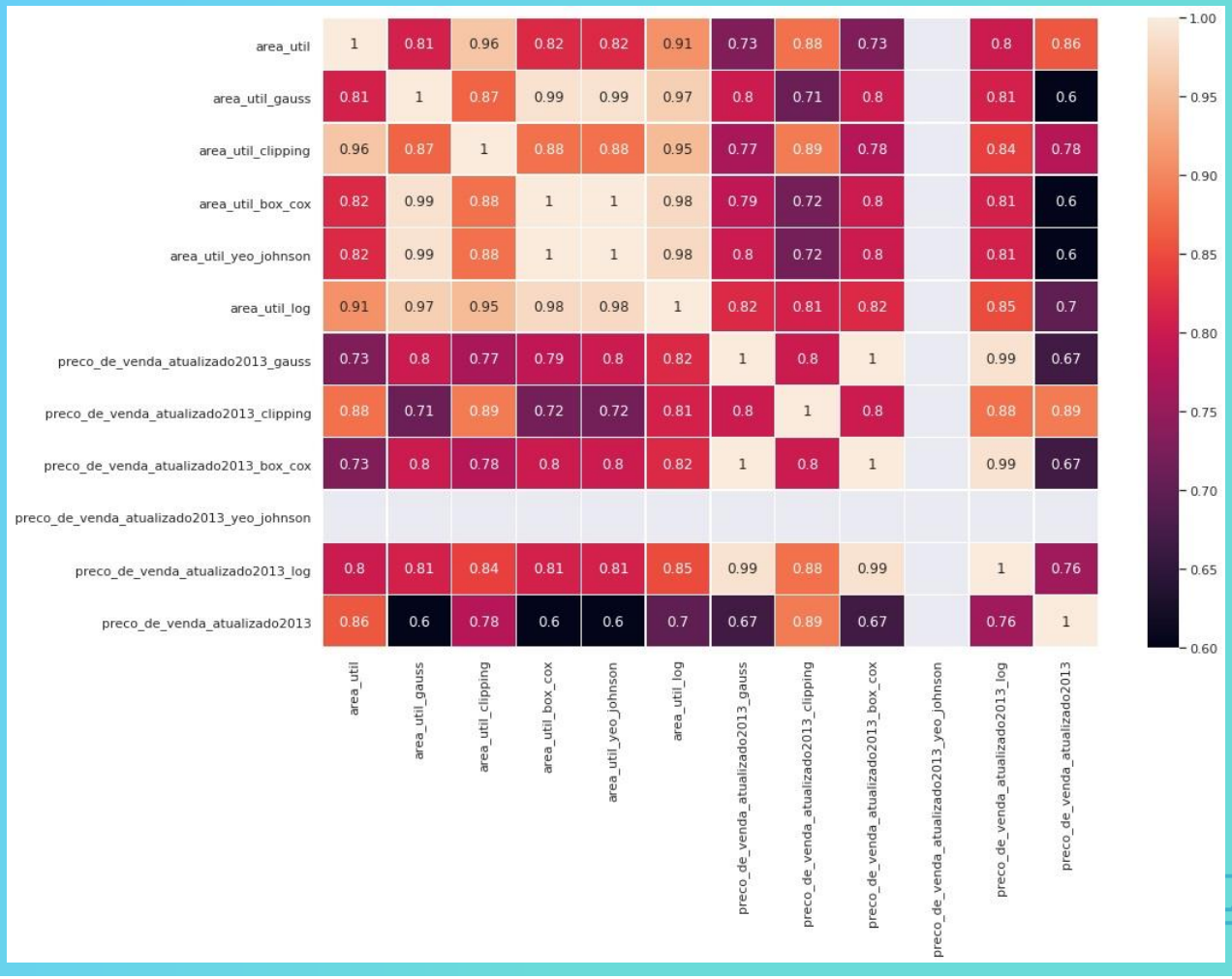
Será utilizada a feature area_util para mostrar o processo que foi seguido por ela e pelas demais variáveis numéricas:

area_util_gauss
area_util_clipping
area_util_box_cox
area_util_yeo_johnson
area_util log

5. Modelagem com estatística

Transformação de variáveis

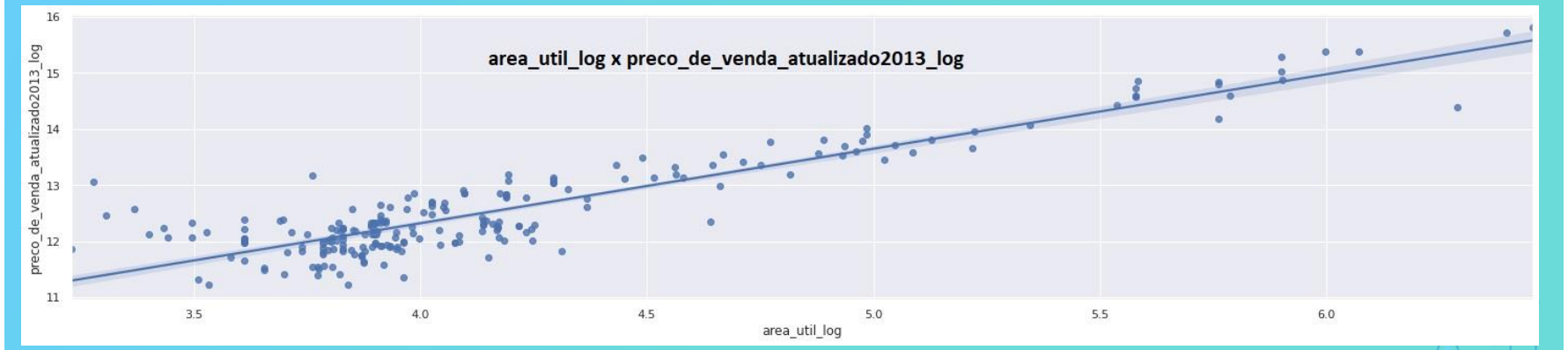
Após a inseção das variáveis na abt trabalhada, no caso da zona 1 é a abt_base1, plota-se a matriz de correlação transformadas e original e verifica-se qual transformação tem melhor correlação com a variável target. Esta transformação é selecionada.



5. Modelagem com estatística

Transformação de variáveis

Após a seleção da melhor feature plotamos o gráfico da variável selecionada em relação a vatiável target para verificar a sua lineridade de forma intuitiva.



Este processo foi feito para as seguintes variáveis numéricas:

- garagens
- area_total
- dormitorios
- banheiros ok
- unidades_por_andar
- andares
- distancia_trem_metro

4. Análise exploratória de dados

E as variáveis categóricas???

5. Modelagem com estatística

Abordagem para variáveis categóricas

- sistema_finaceiro: duas categorias
- setor: 205 categorias
- zona: 54 categorias
- municipio: 25 categorias
- cooperativa: 2 categorias
- hotel: 2 categorias
- flat: 2 categorias
- exflat: 2 categorias
- estacao: 135 categorias

As variáveis categóricas foram tratadas antes da segmentação dos dados

Já abordamos a tratativa da variável zona.
As variáveis setor, municipio e estação foram submetidos ao mesmo processo já descrito para a variável zona e deram origem a variáveis com 3 categorias cada:

estacao_agg: seguiu na análise
municipio_agg: seguiu na análise
setor_agg: por não se mostrar representativo foi eliminado

5. Modelagem com estatística

O Modelo zona1

Foram obtidos três modelo um para cada segmento, observando o modelo criado para o segmanto abt_base1 criada para zona1n usando como target 'preco_de_venda_atualizado2013_log'.

Tabela4: Modelo zona1

feature	coeficiente
constante	7.90
area_util_log	1.26
andares_gauss	0.15
cooperativa_0	0.34
flat_0	-0.86
estacao_agg_grp_estacao1	-0.14

Verifica-se que mantendo-se as demais variáveis constantes:

- o log da area útil eleva o preço do imóvel, como consequencia quanto maior a area util maior o preço do imóvel;
- a variável andares_gauss aumenta o preço do imóvel;
- se o imóvel não for cooperative tem maio valor;
- se o imóvel não for flat tem menor valor;
- Se o imóvel o imóvel estiver mais próximo de alguma estação do grupo 1 ele possuirá menor valor.

5. Modelagem com estatística

O Modelo zona2

Verifica-se que mantendo-se as demais variáveis constantes:

- o log da area útil eleva o preço do imóvel, como consequencia quanto maior a area util maior o preço do imóvel;
- a variável andares_gauss aumenta o preço do imóvel;
- Quanto maior a distancia do imóvel a uma estação de trem ou metrô maior é o decrescimo no seu valor de venda.
- Se for adquirido por preço fechado tem menor preço
- Se não é cooperattiva tem maior preço
- Se não é flat tem menor preço
- Se a estação mais proxima for do grupo estação1 tem menor valor
- Se a estação mais próxima for do grupo estação2 tem maior valor
- Se o imóvel estiver nos grupos municipio1 ou municipio2 ele terá menor valor

Tabela5: Modelo Zona 2

feature	coeficientes
constante	8.50
area_util_log	1.12
andares_gauss	0.04
distancia_trem_metro	-0.03
sistema_financeiro_preco fechado	-0.04
cooperativa_0	0.43
flat_0	-0.70
estacao_agg_grp_estacao1	-0.27
estacao_agg_grp_estacao2	0.44
municipio_agg_grp_municipio1	-0.38
municipio_agg_grp_municipio2	-0.11

5. Modelagem com estatística

O Modelo zona3

Verifica-se que mantendo-se as demais variáveis constantes:

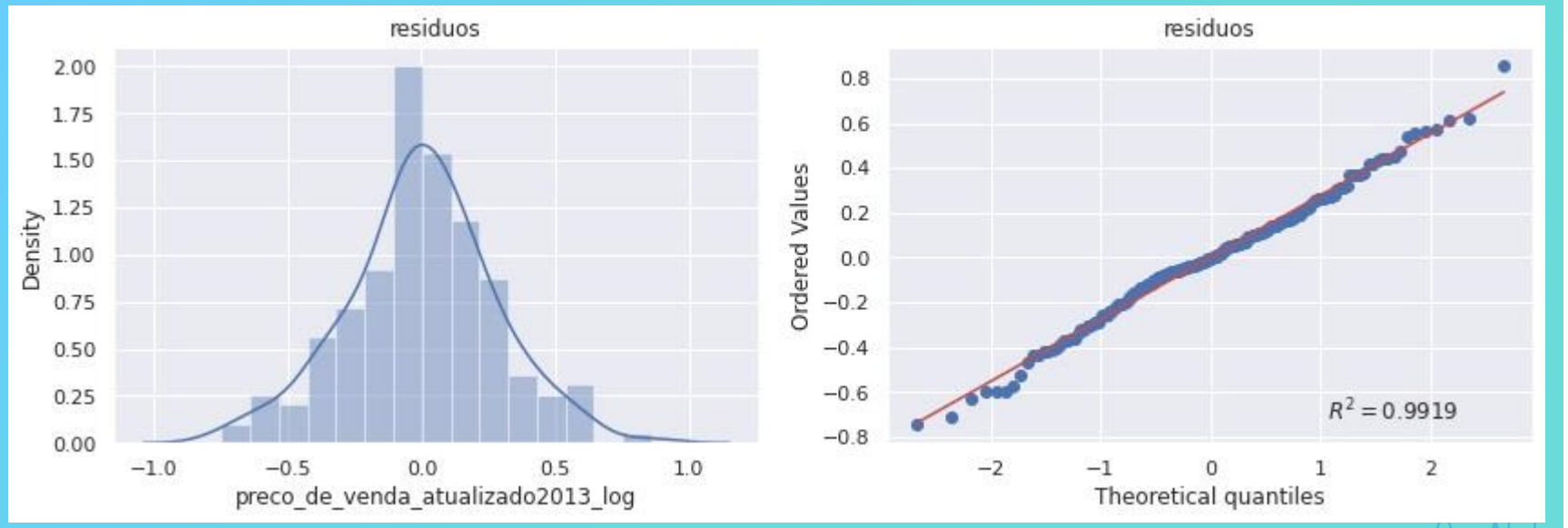
- o log da area útil eleva o preço do imóvel, como consequência quanto maior a area util maior o preço do imóvel;
- Quanto maior a distancia do imóvel a uma estação de trem ou metrô maior é o acrescimo no seu valor de venda.
- a variável andares_gauss eleva o o preço do imóvel;
- Se for adquirido por preço fechado tem maior preço
- Se não é cooperattiva tem maior preço
- Se não é flat tem menor preço
- Se a estação mais proxima for do grupo estação2 ou estação3 tem maior valor

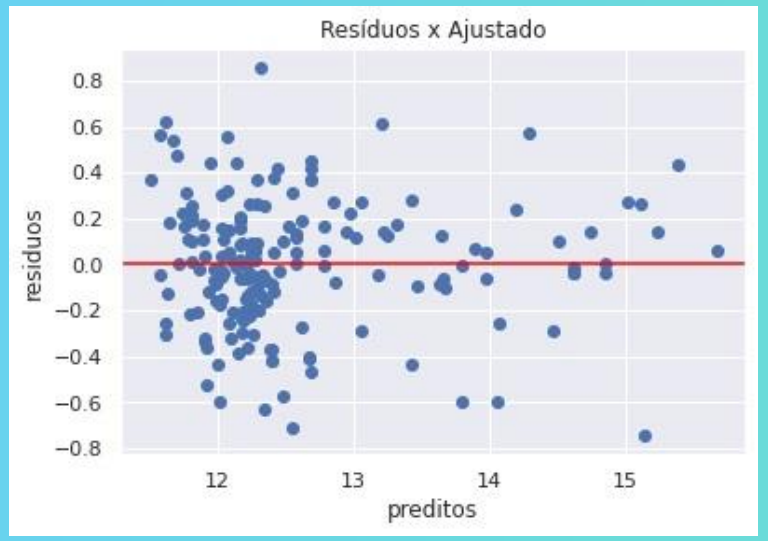
Tabela6: Modelo Zona3

feature	coeficientes
constante	7.69
area_util_log	1.15
distancia_trem_metro	0.02
andares_gauss	0.04
sistema_financeiro_preco fechado	0.08
cooperativa_0	0.34
flat_0	-0.24
estacao_agg_grp_estacao2	0.27
estacao_agg_grp_estacao3	0.48

5. Modelagem com estatística

Análise de Resíduos: teste com 80% da zona1 – criação modelo





Verifica-se que:

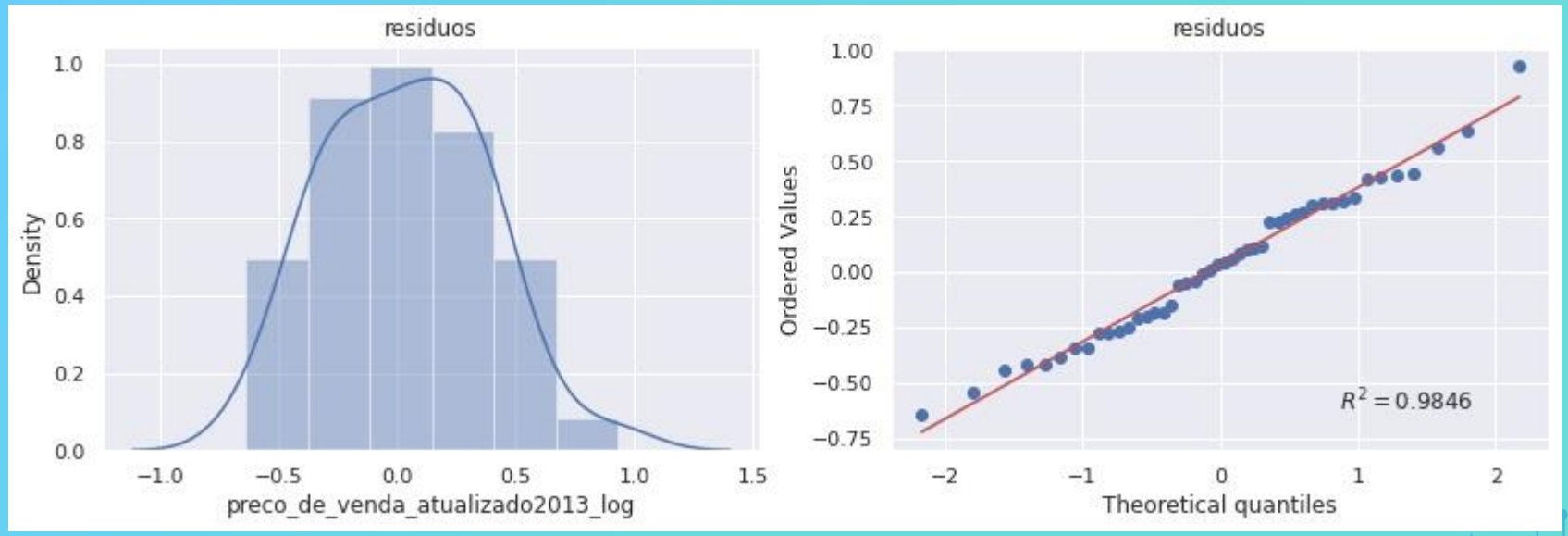
- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

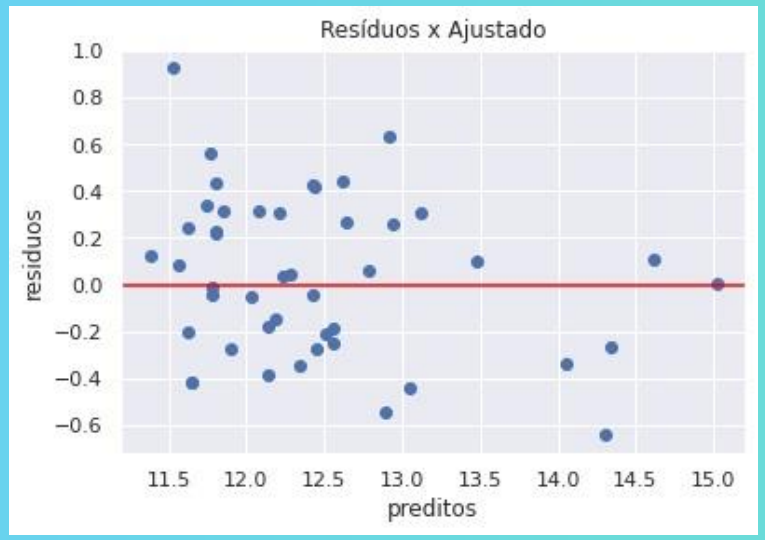
R2-Ajustado= 0.908

MAE: 0.21

5. Modelagem com estatística

Análise de Resíduos: teste com 20% da zona1 – teste modelo





Verifica-se que:

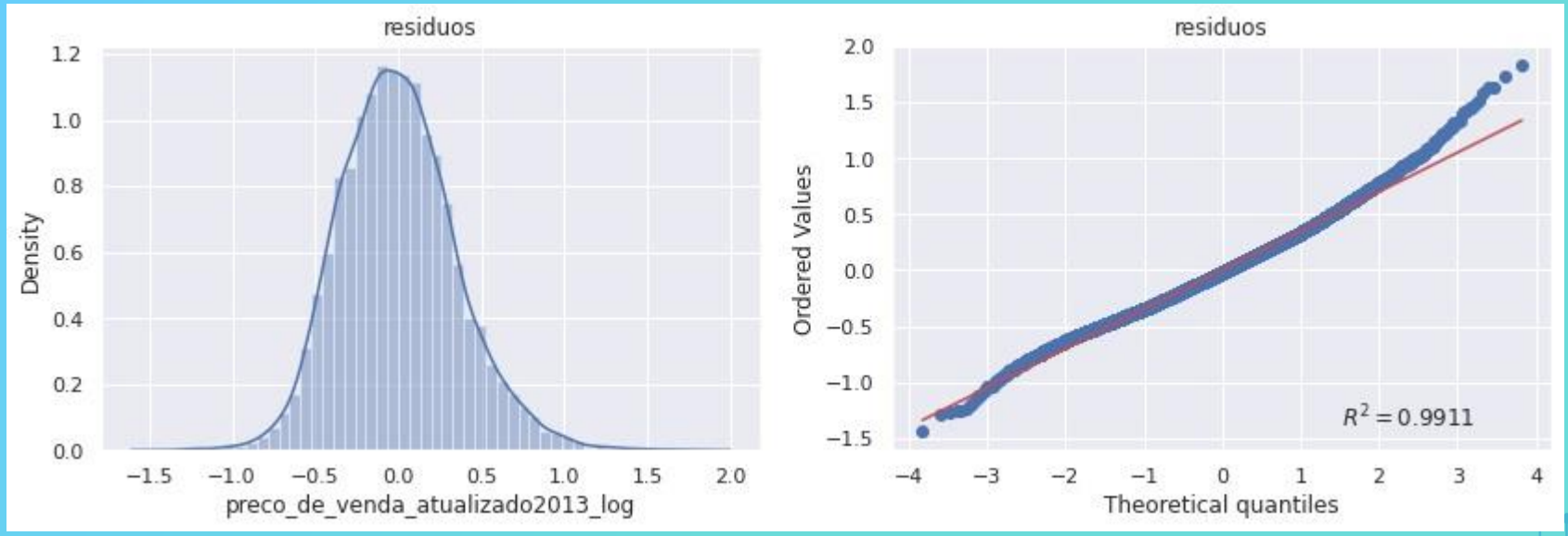
- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

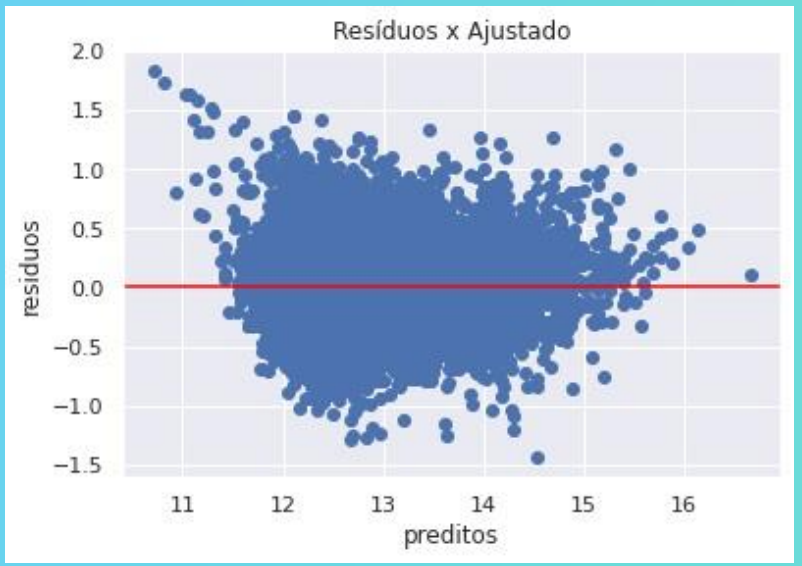
R2-Ajustado= 0.811

MAE: 0.28

5. Modelagem com estatística

Análise de Resíduos: teste com 80% da zona2 – criação modelo





Verifica-se que:

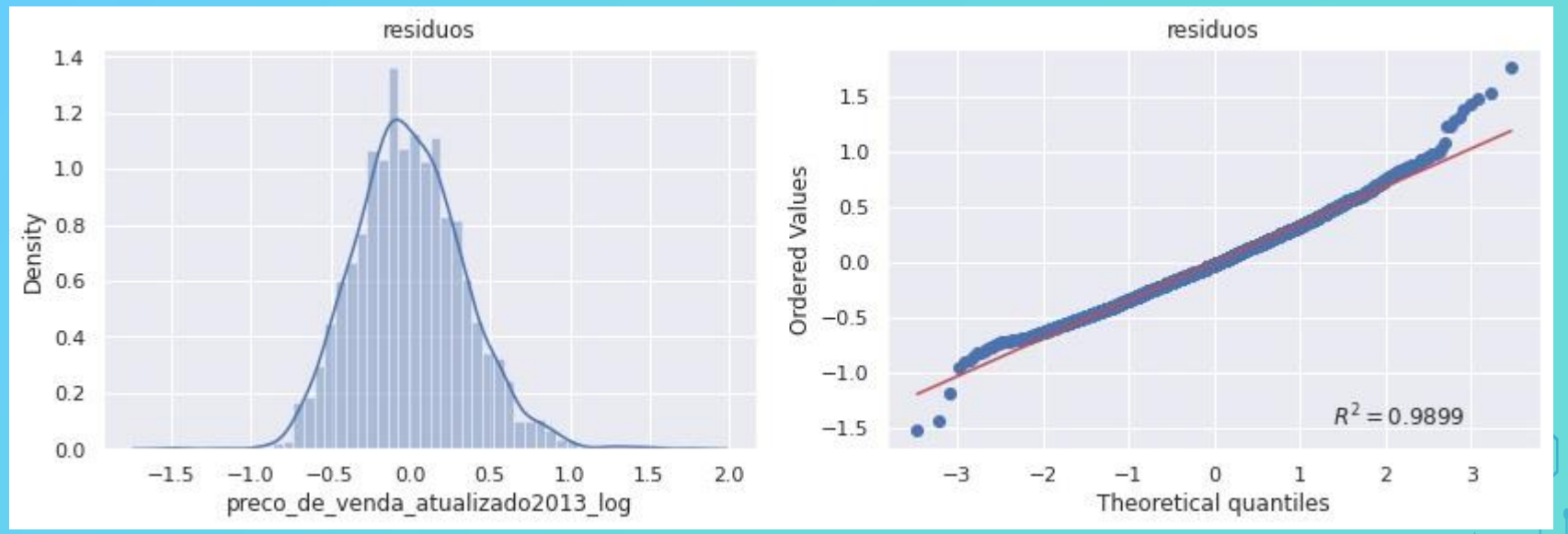
- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

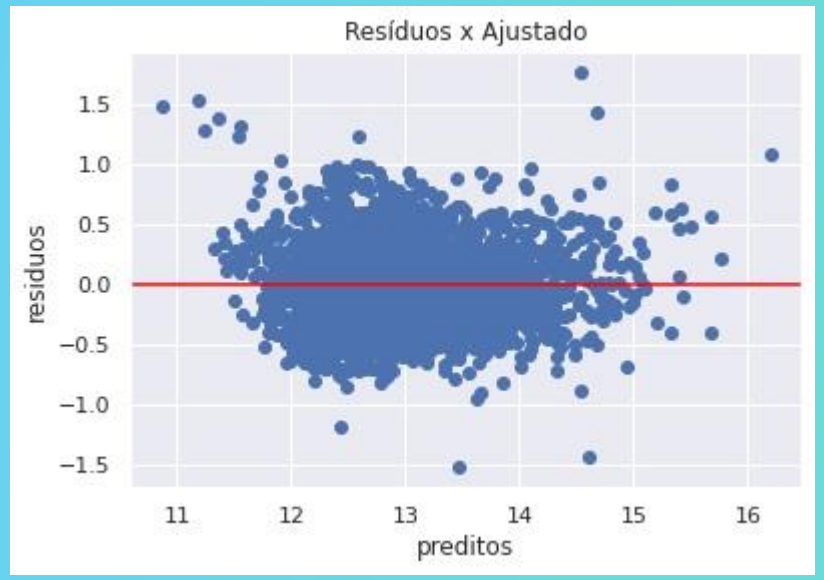
R2-Ajustado= 0.80

MAE: 0.28

5. Modelagem com estatística

Análise de Resíduos: teste com 20% da zona2 – teste modelo





Verifica-se que:

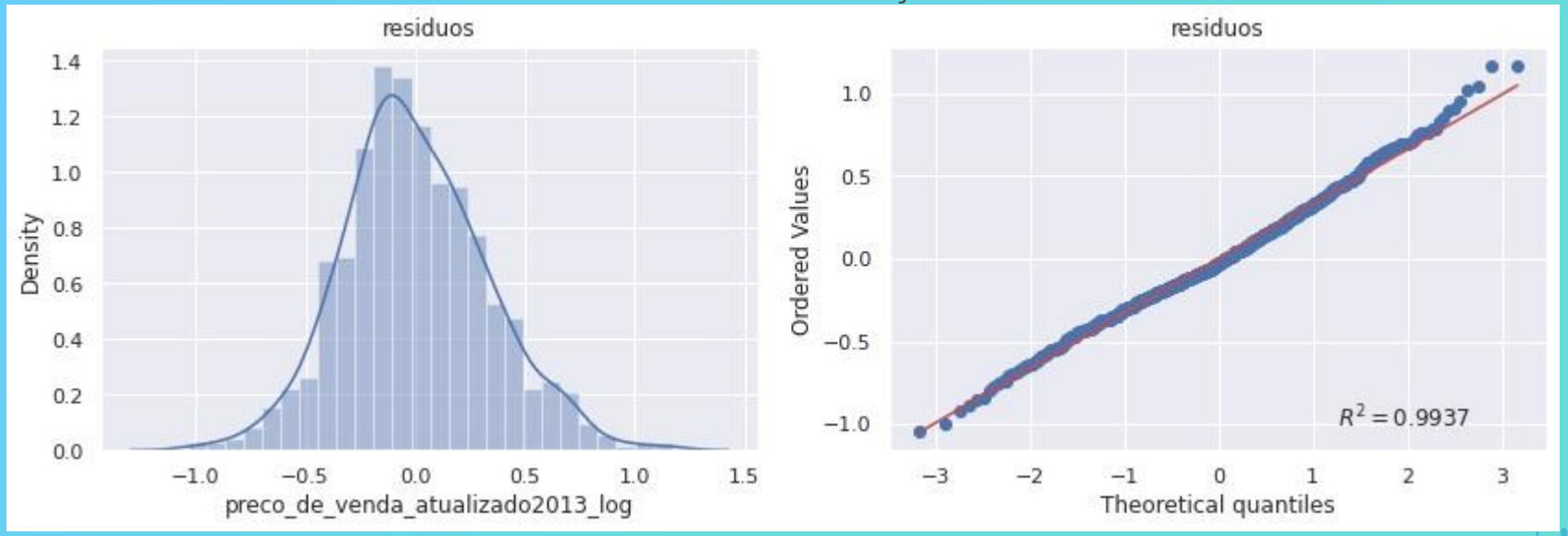
- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

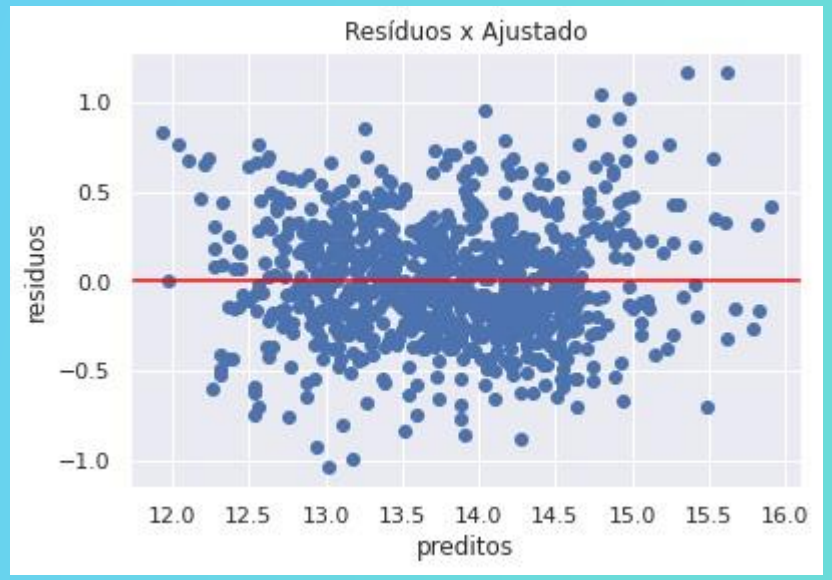
R2-Ajustado= 0.81

MAE: 0.27

5. Modelagem com estatística

Análise de Resíduos: teste com 80% da zona3 – criação modelo





Verifica-se que:

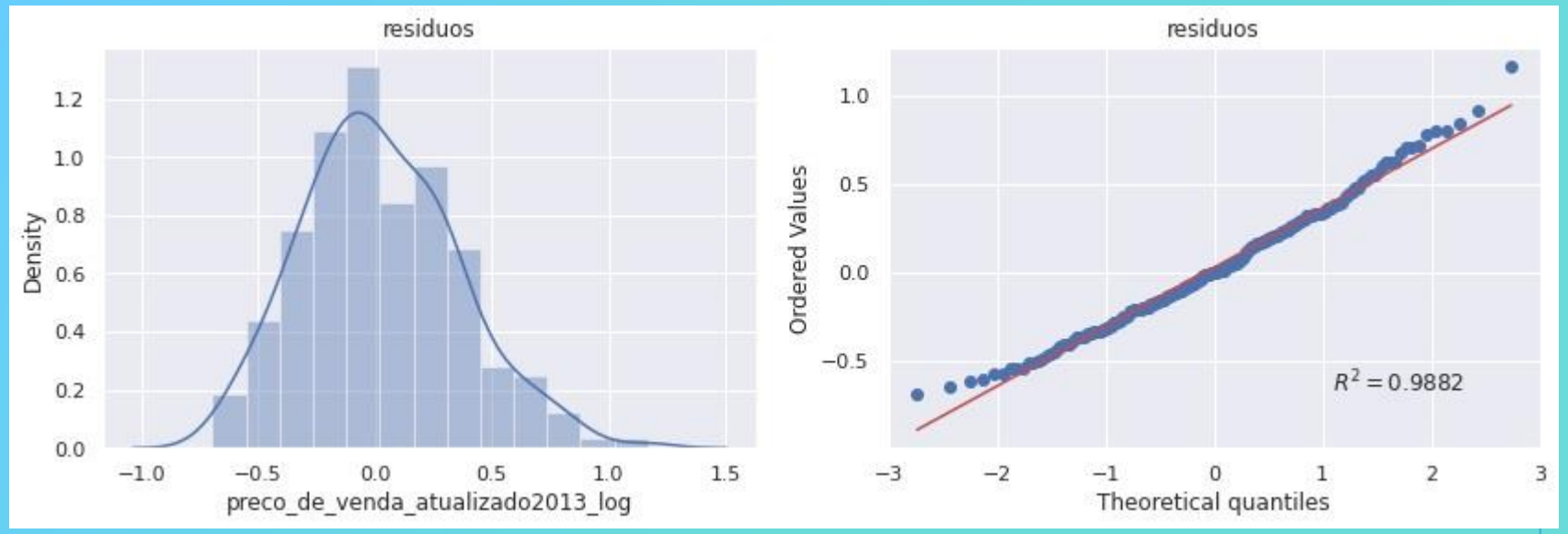
- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

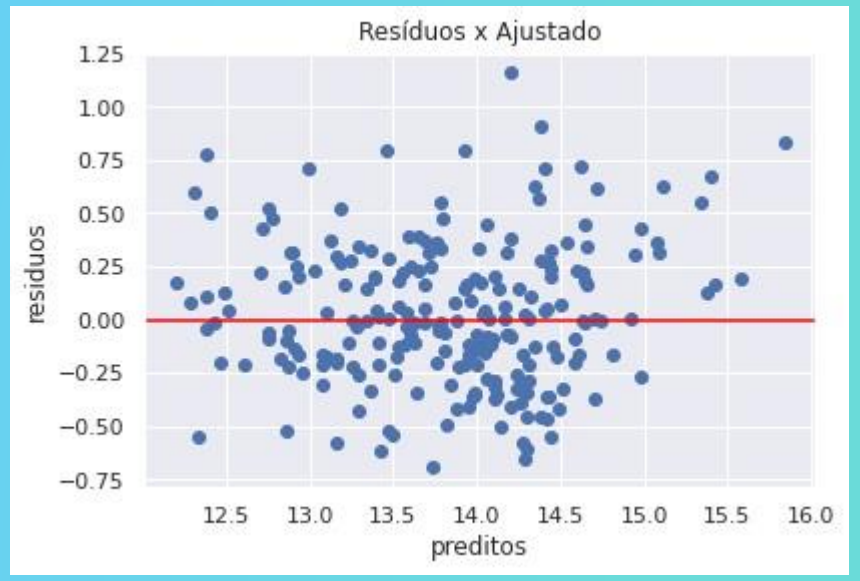
R2-Ajustado= 0.82

MAE: 0.26

5. Modelagem com estatística

Análise de Resíduos: teste com 20% da zona3 – teste modelo





Verifica-se que:

- Os resíduos estão distribuídos de forma simétrica em torno de 0.
- Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica

R2-Ajustado= 0.82

MAE: 0.27

5. Modelagem com inteligência Artificial

Abordagem:

- Para iniciar esta modelagem serão utilizadas as variáveis já descritas no início da modelagem estatistica.
- Efetivamente serão utilizados as segmenações abt_base1, abt_base2 e abt_base3 contudo será aplicada a transformação de variável unicamente para a variável target.
- A transformação será logaritmica conforme foi aplicada na modelagem estatística.

Foram aplicados os seguintes modelos a cada segment de dados:

- Regressão Linear
- Ridge regression
- Decision Tree Regression
- Randon Forest Regression
- LGBM Regression
- XGBoost Regression
- CatBoost Regression

Os modelos serão avaliados e considerando as seguintes métricas o melhor modelo será seecionado e depois otimizado:

- R2-ajustado
- mse

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

6. Modelagem com inteligência Artificial

Modelo Selecionado para Zona1: CatBoost Regression

A modelagem com inteligência artificial apresentou melhores resultados.

Área util e área total foram as features mais importantes para o modelo.

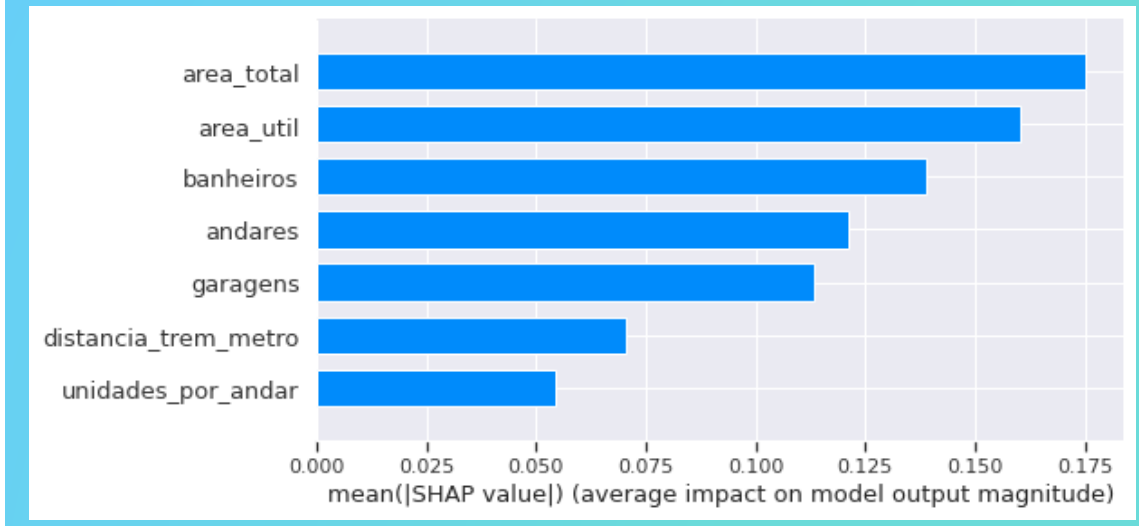
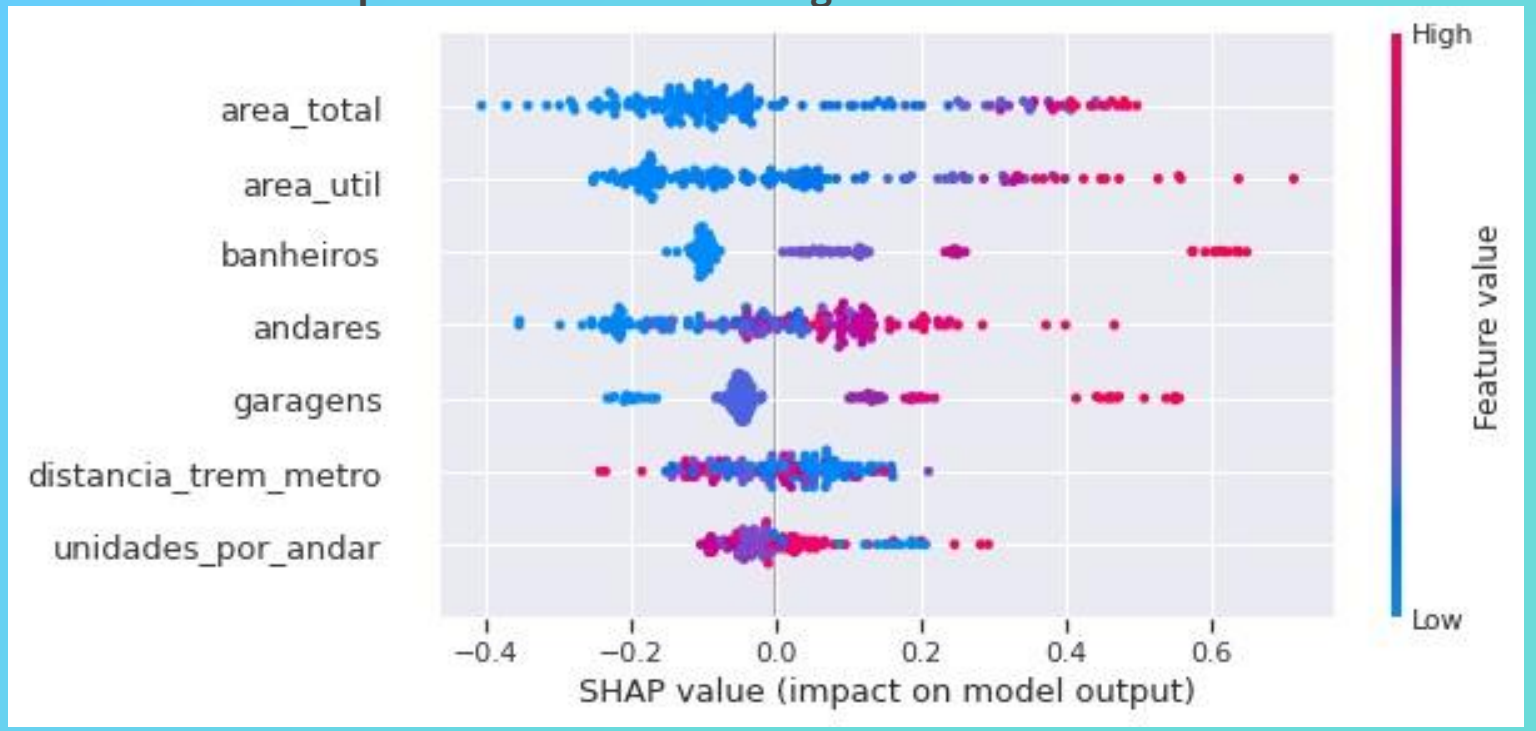


Tabela7: Métricas

métrica	treino	teste
r2_ajustado	0.998	0.932
mse	0.001	0.057

6. Modelagem com inteligência Artificial

Modelo Selecionado para Zona1: CatBoost Regression



- Quanto maior a area total, maior o preço do imóvel
- Quanto maior a área útil, maior o preço de venda do imóvel
- Quanto mais banheiros maior o preço do imóvel
- Quanto menor a distância das estações de metro ou trem maior o preço do imóvel
- Quanto menos unidades houverem por andar maior o preço do imóvel

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

6. Modelagem com inteligência Artificial

Modelo Selecionado para Zona2: CatBoost Regression

A modelagem com inteligência artificial apresentou melhores resultados.

Área util e área total foram as features mais importantes para o modelo.

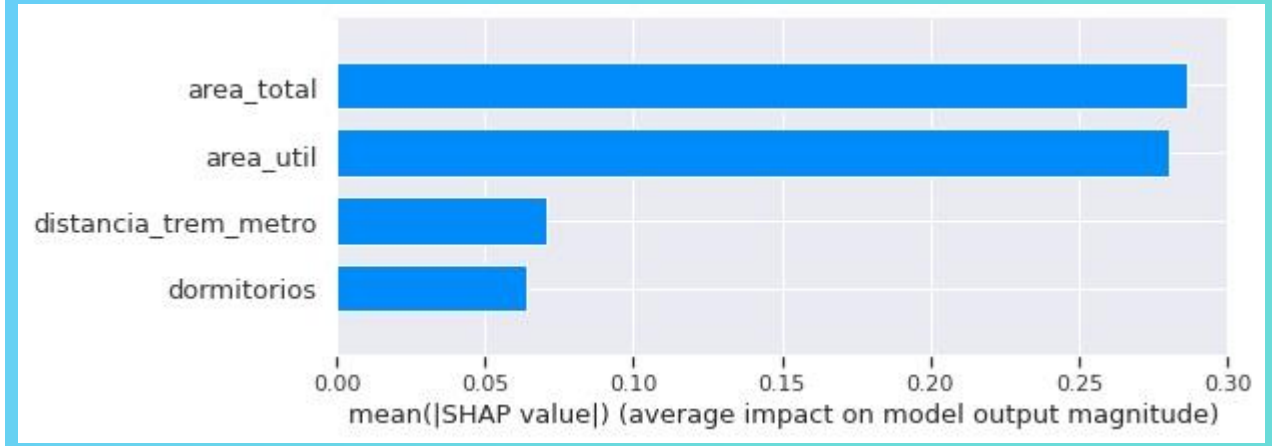
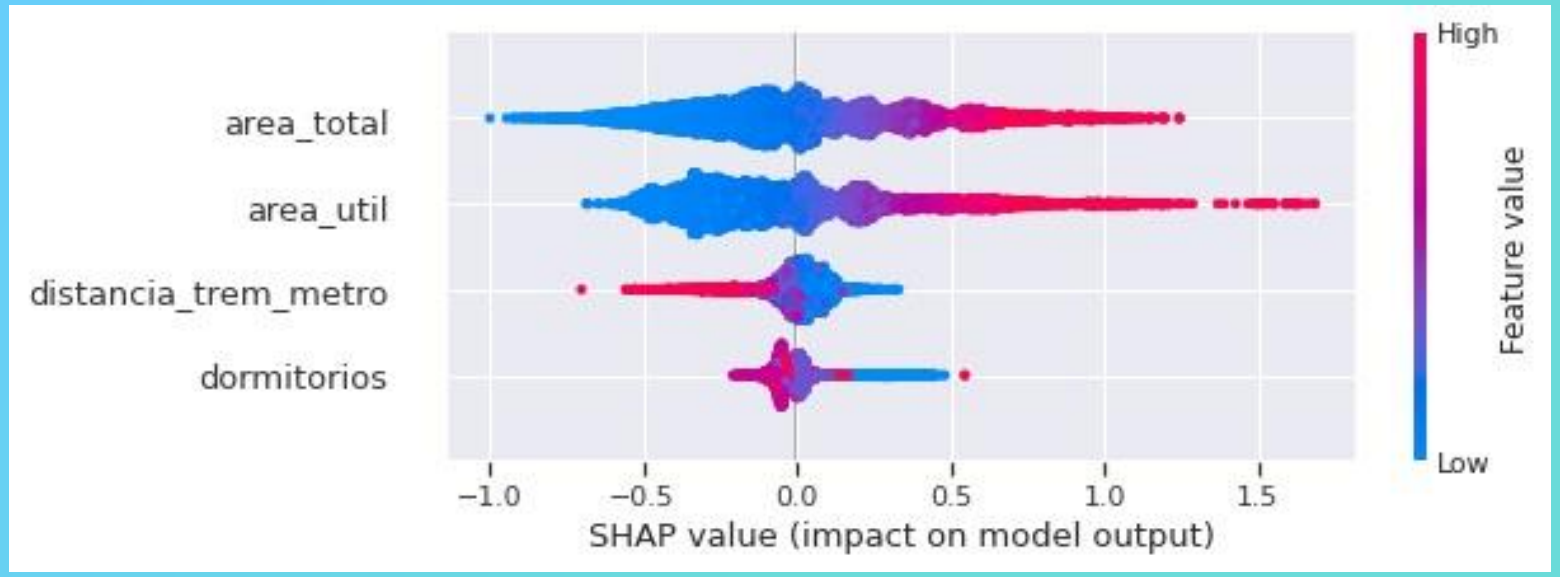


Tabela8: Métricas

métrica	treino	teste
r2_ajustado	0.900	0.864
mse	0.062	0.081

5. Modelagem com inteligência Artificial

Modelo Selecionado para Zona2: CatBoost Regression



- Quanto maior a area total, maior o preço do imóvel
- Quanto maior a área útil, maior o preço de venda do imóvel
- Quanto menor a distância das estações de metrô ou trem maior o preço do imóvel
- Quanto menos dormitórios houverem maior o preço de venda do imóvel

6. Modelagem com inteligência Artificial

Modelo Selecionado para Zona3: LGBM Regression

A modelagem com inteligência artificial apresentou melhores resultados.

Área util e área total foram as features mais importantes para o modelo.

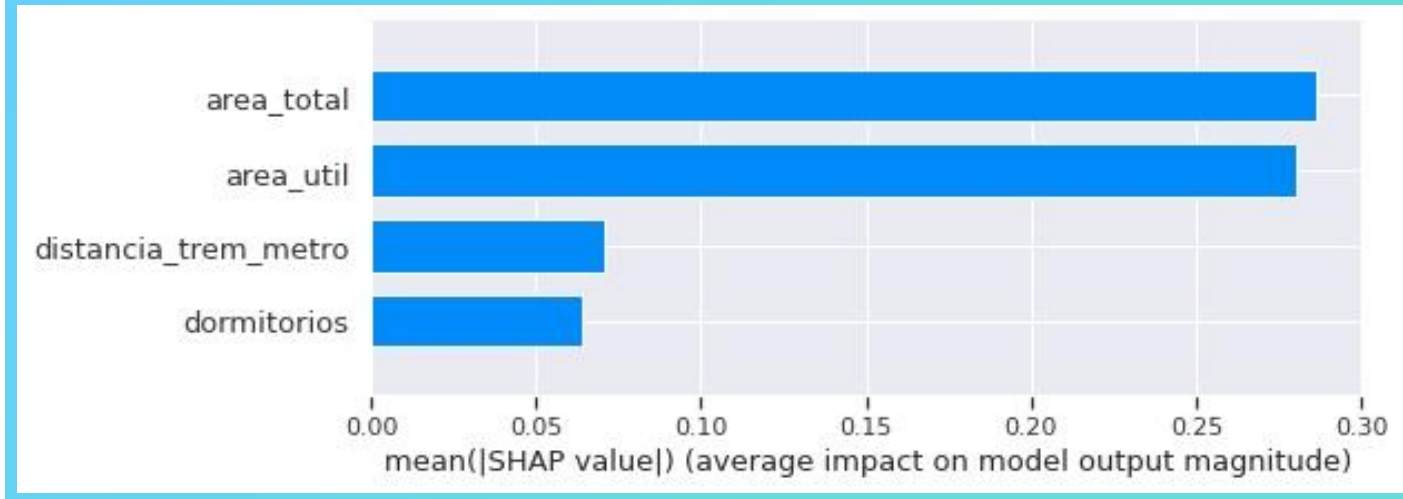
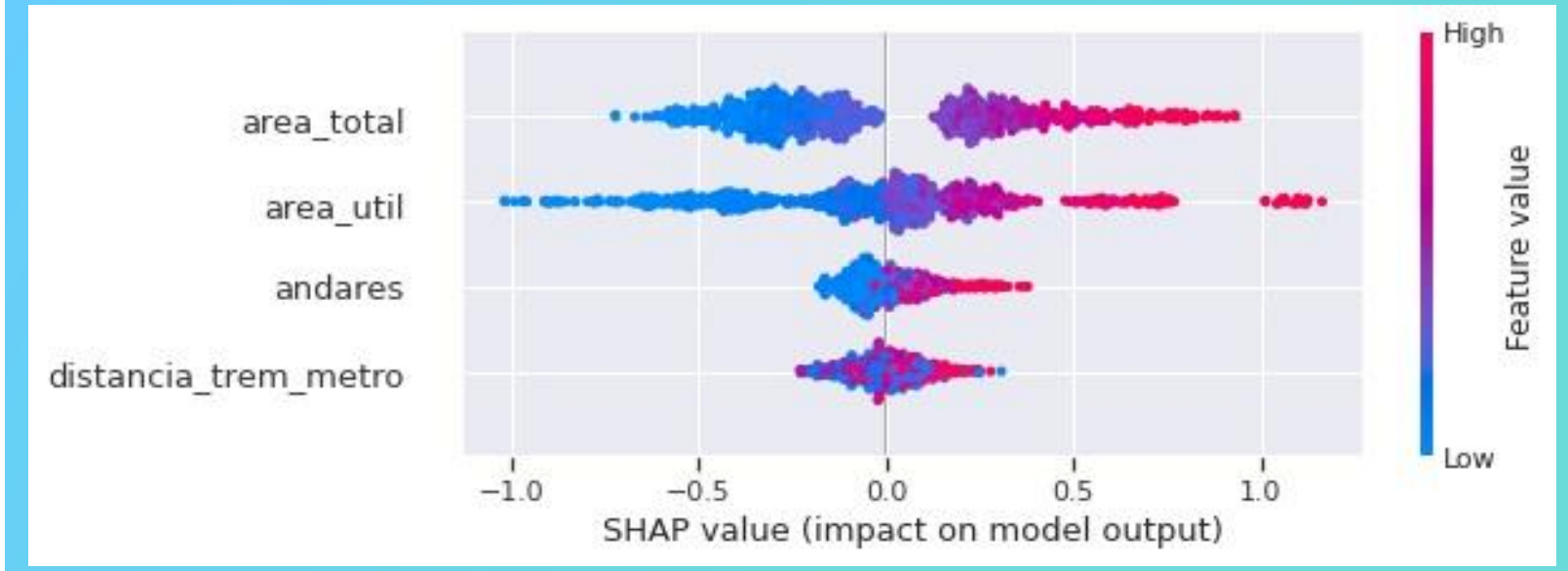


Tabela9: Métricas

métrica	treino	teste
r2_ajustado	0.944	0.860
mse	0.022	0.097

6. Modelagem com inteligência Artificial

Modelo Selecionado para Zona2: LGBM Regression



- Quanto maior a area total, maior o preço do imóvel
- Quanto maior a área útil, maior o preço de venda do imóvel
- Quanto mais andares maior o preço do imóvel
- Quanto maior a distância de estações de trem ou metrô maior é o preço do imóvel

7. Conclusões

Para a criação de um modelo para estes dados a principal dificuldade foi encontar uma abordagem que conseguisse diminuir a quantidade de outliers, as segmentações tradicionais não mostraram-se eficientes.

A abordagem utilizada foi definida por tentative e erro.

A escolha da transformação logarítmica foi pensada porque o modelo em questão não apresentava desvio padrão aproximadamente constante.

Olhando para os modelos criados podemo-se reforcar a importancia das features área útil e área total para a definição do modelo, assim como a variável criada distância trem metrô mostrou-se bastante útil para a modelagem via inteligência artificial.

Acredito que ficam duas conclusões principais:

A importância da análise exploratória e preparação dos dados antes da aplicação do modelo em sí.

A importância do conhecimento do negócio para qual os dados estão sendo modelados é de grande importância para o sucesso do modelo pois problemas reais apresentam complexidade intinseca do próprio negócio.

8. Sugestões para trabalhos futuros

Uma sugestão para trabalhos futuros seria inserir novas features ao modelo e utilizar redes neurais para fazer a análise.