CESAR SCHOOL
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RODRIGO ANTONIO LIRA DE ALBUQUERQUE

USO DE APRENDIZADO DE MÁQUINA PARA OTIMIZAR AS DECISÕES RELATIVAS A INVESTIMENTOS IMOBILIÁRIOS RESIDENCIAIS

Recife,
2024

RODRIGO ANTONIO LIRA DE ALBUQUERQUE

USO DE APRENDIZADO DE MÁQUINA PARA OTIMIZAR AS DECISÕES RELATIVAS A INVESTIMENTOS IMOBILIÁRIOS RESIDENCIAIS

Trabalho de conclusão de curso apresentado ao Programa de Graduação em Ciência da Computação da CESAR SCHOOL, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Anderson Tenorio Sergio
Coorientador: Prof. MSc. Eronides Felisberto da Silva Neto.

Recife
2024

AGRADECIMENTOS

Agradeço ao meu orientador, Anderson, por todo o suporte e auxilio ao longo deste trabalho, ao meu coorientador Eronides, por ter me acompanhado durante todo o início dessa trajetória, sendo peça fundamental em diversas ideias e dicas para elaboração desta monografia. Meu reconhecimento vai também aos professores da graduação, que se dedicaram com paixão a trazer o que há de mais moderno no mercado, enriquecendo minha formação. Além disso, não poderia deixar de agradecer aos meus colegas, que transformaram esta jornada de uma segunda graduação – que tinha tudo para ser desafiadora e exaustiva – em uma experiência muito mais leve, enriquecedora e gratificante.

Por último, agradeço principalmente aos meus pais. Eles foram fundamentais, lembro, até hoje, dos momentos em que meu pai chegava cansado do trabalho e parava tudo para me ensinar a ler, ou da minha mãe, que se desdobrava para cuidar de mim. Sem eles, eu não seria nada neste mundo. Não sou muito bom com palavras, mas com certeza este trabalho é um reflexo da contribuição de todos vocês.

RESUMO

Os imóveis representam um grande negócio, seja para uso próprio ou como investimento. No entanto, as decisões relacionadas a esse tipo de compra ainda são fortemente ligadas à emoção, o que evidencia a necessidade de métodos mais racionais de decisão. Neste trabalho, busca-se implementar um modelo de aprendizado de máquina que tente prever o preço justo de um imóvel. A ideia é desenvolver uma estrutura que seja fácil de aplicar e, de preferência, utilize técnicas amplamente disponíveis na literatura.

Para a modelagem, dados brutos foram colhidos, vindos de uma grande plataforma imobiliária online. O escopo escolhido foram apartamentos, com custo abaixo de R$ 2 milhões e localizados em 11 bairros de classe média alta da cidade de São Paulo. Apesar do escopo ser reduzido à cidade de São Paulo, os passos elencados aqui para lidar com os dados podem ser facilmente replicados para qualquer localidade, com pequenos ajustes.

Retirar dados de *websites* imobiliários vem acompanhado de inúmeros desafios. Frequentemente, o preço anunciado diverge do preço efetivo de venda, os dados podem estar preenchidos de forma incorreta, e há também casos de valores muito acima ou abaixo da média de mercado. Encontrar uma maneira de suavizar esses valores discrepantes é de extrema importância numa modelagem imobiliária.

Além disso, os dados brutos obtidos nem sempre estão prontos para o uso imediato. Antes mesmo de começar uma modelagem, há desafios significativos para a preparação e transformação da base. O pipeline de tratamento é essencial. O trabalho aqui apresentado inclui técnicas de pré-processamento, limpeza, estratégias para lidar com dados ausentes, transformações, feature engineering, normalização dos dados, codificação, redução de dimensionalidade (como o algoritmo PCA), além de abordagens para tratar *outliers* e selecionar as características principais, como o *Recursive Feature Elimination* (RFE).

Para evitar o *overfitting* também foram utilizados métodos de divisão da base de dados como o *Holdout*, além de validação cruzada *K-fold*. No fim, para estimar os preços, dois métodos clássicos de aprendizado de máquina foram utilizados: *random forest* e análise de regressão. O método baseado em árvore de decisão foi o que apresentou o desempenho mais satisfatório.

Palavras-chave: Imóveis; Precificação de Apartamentos; Random Forest; Aprendizado de Máquina aplicado a investimentos imobiliários; Análise de Regressão.

ABSTRACT

Real estate represents a significant business opportunity, whether for personal use or as an investment. However, decisions related to this type of purchase are still heavily influenced by emotions, highlighting the need for more rational decision-making methods. This study aims to implement a machine learning model to predict the fair price of a property. The goal is to develop a structure that is easy to apply and, preferably, utilizes techniques widely available in the literature.

For the modeling, raw data was collected from a large online real estate platform. The chosen scope included apartments costing less than R$ 2 million and located in 11 upper-middle-class neighborhoods in the city of São Paulo. Although the scope is limited to São Paulo, the steps outlined here to handle the data can be easily replicated for any location, with minor adjustments.

Extracting data from real estate websites comes with numerous challenges. Frequently, the listed price differs from the effective sale price, the data may be incorrectly filled in, and there are also cases of values far above or below the market average. Finding a way to smooth out these outliers is of utmost importance in real estate modeling.

Moreover, the raw data obtained is not always ready for immediate use. Even before starting modeling, there are significant challenges in preparing and transforming the dataset. The treatment pipeline is essential. The work presented here includes preprocessing techniques, cleaning, strategies for handling missing data, transformations, feature engineering, data normalization, encoding, dimensionality reduction (such as the PCA algorithm), as well as approaches for dealing with outliers and selecting the main features, such as Recursive Feature Elimination (RFE).

To avoid overfitting, methods such as Holdout and K-fold cross-validation were also used. In the end, to estimate prices, two classic machine learning methods were employed: random forest and regression analysis. The decision-tree-based method achieved the most satisfactory performance.

Keywords: Real Estate; Apartment Pricing; Random Forest; Machine Learning Applied to Real Estate Investments; Regression Analysis.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1. Introdução

1.1 Motivação e Contexto

O mercado imobiliário residencial brasileiro é um dos pilares que influencia diretamente a economia do país. Além de moradia e infraestrutura, o aquecimento do setor impacta outras áreas como, por exemplo, a construção civil. Pode-se ressaltar também que, a partir do momento que surgem novas residências, inúmeros comércios e serviços acessórios também aparecem no entorno para oferecer suporte. A perspectiva de crescimento para os próximos anos do setor é animadora.

Segundo um relatório da Mordor Intelligence (2023), uma empresa global especializada em análise de mercado, o tamanho desse mercado, em solo nacional, equivale à cerca de US$ 59,61 bilhões em 2024 e tenderá a crescer para para até US$ 77,54 bilhões até 2029. Além disso, dados divulgados em 2024 num relatório emitido pela Câmara Brasileira da Indústria da Construção (CBIC), relativos à análise do quarto trimestre de 2023, mostrou que programas, como o Minha Casa Minha Vida, podem servir como uma forma de impulsionamento para o setor. A previsão de crescimento para o ano de 2024 gira em torno de 5% a 10% para imóveis relacionados ao Minha Casa, Minha Vida e entre 0% a 5% para os demais imóveis. (CBIC, 2024)

Muito além da função primária de moradia, os imóveis também se destacam como uma opção válida para diversificar o patrimônio. Além dos tradicionais investimentos em renda fixa, como os do Tesouro Direto, ou dos em renda variável, como ações, os imóveis oferecem uma alternativa extra para diversificar o portfólio. Uma prática bastante comum é adquirir imóveis com o objetivo de alugá-los, seja a longo prazo, como tradicionalmente sempre se fez, ou ainda se aproveitar do crescimento de plataformas imobiliárias, como o Airbnb, que simplificam o processo de locação e oferecem maior flexibilidade ao proprietário, permitindo aluguéis de curto prazo.Toda essa flexibilidade evidencia a atratividade do setor imobiliário.

O uso do aprendizado de máquina aplicado ao setor de imóveis ainda tem muito espaço para expansão, bons trabalhos no contexto brasileiro foram elaborados recentemente como o de Sousa (2023), que utilizou *Random Forest* para estimar valores de imóveis em Aracaju-SE ou o de Miranda et al. (2023), que combinou *Random Forest*, *Support Vector Machines* (SVM) e Regressão Linear com

foco em estimar preços de aluguel. Apesar de existir uma variedade crescente de bons trabalhos recentes na literatura nacional, o número ainda é pouco comparado às inúmeras possibilidades de arranjo do setor e, atualmente, a maioria das decisões de compra e venda de um imóvel ainda são tomadas com base na emoção e não na racionalidade.

1.2 Problema de Pesquisa

A busca é por implementar um modelo de aprendizado de máquina que seja fácil de implementar e que tenha uma boa estimativa de predição de preços, considerando um comparativo entre o valor real do imóvel e o valor predito pelo modelo. O objetivo é conseguir apontar o valor justo de mercado de um imóvel, para com isso perceber dentro de uma lista de imóveis quais aqueles que apresentam preços tidos como “barganha”, abaixo da média de preço do mercado, se mostrando boas oportunidades de compra. Com essa lista de principais imóveis potenciais, o utilizador do modelo consegue filtrar e direcionar seu escasso tempo para fazer visitas presenciais mais certeiras, em imóveis com bom potencial em termos de retorno de investimento.

As possibilidades de pesquisa são inúmeras dada a variedade do tema. Esse presente trabalho restringe o escopo para imóveis residenciais, a fim de ter uma lógica que seja útil tanto para grande quanto para pequenos investidores, ou até mesmo para quem deseja ter uma alternativa racional de compra para seu único imóvel de moradia.

A base da pesquisa se baseia em usar uma ferramenta de *scraping* para extrair dados das principais plataformas de venda imobiliária nacionais. Após adquirir esses dados brutos, tratá-los, e selecionar um modelo de aprendizado de máquina que mais se mostre adequado para analisá-los com base nas informações dos imóveis. Assim, é possível perceber quais imóveis aparentam estar com preços justos, além de buscar extrair *insights* que possam vir a auxiliar na tomada de decisão.

A análise precisa ser feita com cuidado, pois o setor imobiliário apresenta muitos desafios e peculiaridades. Cada localidade pode ter fatores completamente distintos impactando no preço dos imóveis. Por isso, não é incomum encontrar modelos que desempenhem bem em uma região, e não tão bem em outras. Dessa

forma, a generalização tem que ser feita de forma cautelosa. Por exemplo, para uma certa cidade litorânea um dos fatores mais importantes pode ser a distância do imóvel para a praia. Já numa cidade sem praia, a distância para parques ou museus pode ser levada em consideração. É possível até que numa mesma cidade existam bairros com características distintas que impactem no preço dos imóveis.

Além disso, é possível que num mesmo prédio, com tamanho e plantas similares, dentro de um mesmo condomínio e localidade, possam existir apartamentos com preços completamente distintos. Isso pode acontecer devido ao estado de conservação do imóvel, reformas, ou outros fatores de difícil quantificação. Outro desafio reside no fato de que nem sempre o preço anunciado é, de fato, o valor a ser vendido no final. Os desafios da pesquisa são inúmeros, mas estar cientes deles ajudará a minimizá-los.

PROBLEMA DE PESQUISA:

> Como estabelecer um modelo de aprendizado de máquina para otimizar as decisões relativas a investimentos imobiliários residenciais, considerando a boa performance na predição de preços e a facilidade de implementação, a fim de tornar as decisões de compra de um imóvel um ato mais racional, promovendo uma estimativa de preço que possibilite uma tomada de decisão mais assertiva ao interessado?

1.3 Justificativa

Comprar um imóvel é algo que faz parte da vida de um número significativo de pessoas em algum ponto de suas trajetórias, marcando um capítulo significativo na construção de cada história pessoal. Seja para uso pessoal, ou como investimento, o ato de comprar ou vender um imóvel é sempre difícil. Mesmo sendo dotado de tamanha importância e envolvendo às vezes economias financeiras de uma vida, a maioria das decisões sobre compra ou venda de imóveis ainda são feitas de forma majoritariamente emocional, sem um sistema racional e lógico auxiliando na tomada de decisão. Um estudo divulgado na revista Exame por Quesada (2024) aponta que 84% dos entrevistados têm dificuldade de estimar o preço adequado de um imóvel.

É comum para pessoas que querem comprar um imóvel, e nunca o fizeram, se assustarem inicialmente com tamanha variedade de opções e preços. Sabe-se que na presente sociedade, o tempo das pessoas é escasso e fica difícil decidir, no meio de um tremendo leque de opções, que imóveis vale a pena uma visita ou não. O presente trabalho traz um importante benefício ao mostrar aqueles imóveis que possivelmente estão com preços vantajosos, por estarem com valores abaixo do justo. Com isso, o interessado consegue ser mais assertivo na seleção de suas visitas.

A utilidade não se restringe apenas para quem quer comprar um imóvel, mas também para os que desejam uma venda. Muitas vezes é difícil saber quanto cobrar no valor de venda no imóvel que será anunciado. O preço precisa ser compatível com o mercado, ser atrativo para venda. Porém, não é algo trivial de estimar, principalmente para pessoas sem experiência no setor imobiliário. Um valor baixo demais implica em perda financeira. Por outro lado, um valor alto demais pode fazer com que o imóvel demore a ser vendido, acarretando em custos fixos altos com condomínio, taxas, impostos, que pesam no bolso de quem quer vender. Portanto, uma correta precificação é fundamental, tanto olhando pela ótica do comprador, quanto para o vendedor.

O modelo além de ser útil para pessoas físicas, também pode vir a auxiliar pessoas jurídicas, como bancos, fundos de investimento e corretoras imobiliárias oferecendo uma alternativa aos seus atuais processos avaliativos. O *output* do modelo oferecerá uma sugestão de preço que estará dentro da média do mercado, dada as características do imóvel específico.

Decidiu-se utilizar um *scrapper* para extrair os dados devido a dificuldade de encontrar bons conjuntos de dados nacionais, com informações atualizadas. Algumas prefeituras disponibilizam o valor de venda apartamento, mas o conjunto de dados, geralmente, é limitado e se restringe ao montante financeiro e não considera as inúmeras outras características dos imóveis que podem vir a influenciar no valor final, perdendo, assim, informações importantes.

Por isso, decidiu-se extrair os dados de uma plataforma imobiliária, visando simular ao máximo a experiência que alguém que vai comprar um imóvel usualmente se depara. A ideia é ter um modelo que sirva como um guia comparativo para o decisor. Seja para avaliar se existe uma boa oportunidade para a

compra do imóvel, ou até mesmo para oferecer um preço base de referência para montar uma boa estratégia de venda.

1.4 Objetivos

1.4.1 Objetivo geral

Implementar um modelo de aprendizado de máquina para previsão de preços imobiliários, focando em encontrar o seu valor justo de mercado e na sua facilidade de implementação.

1.4.2 Objetivos específicos

- OE1: Identificar as técnicas de aprendizado de máquina, amplamente divulgadas pela literatura, que sejam adequadas para tratar os dados e modelar o problema.
- OE2: Integrar os dados obtidos pelo *scrapper*, tratá-los adequadamente e desenvolver estratégias para lidar com possíveis inconsistências de valores presentes nesse conjunto de dados.
- OE3: Avaliar o modelo preditivo por meio de métricas de desempenho, buscando alcançar um equilíbrio entre precisão e erro.

2. Fundamentação Teórica

Avaliar imóveis é uma tarefa complexa e repleta de nuances. Por isso, existe uma norma brasileira focada exclusivamente em avaliações de bens e que serviu como orientação em algumas partes deste trabalho: a NBR 14653-2 de 2011. Segundo a norma, um dos métodos para identificar o valor de um bem é o método comparativo direto de dados de mercado. Esse método se baseia na comparação entre o bem que se deseja estimar o valor com imóveis semelhantes negociados no mercado num intervalo temporal recente. Por isso, se buscou uma ferramenta de *scraping* que retirasse de sites imobiliários imóveis similares.

2.1 Pré-Processamento de Dados

O pré-processamento de dados é uma das fases mais críticas do trabalho. Os dados, retirados utilizando uma API de *scraping*, no geral, vem de forma desorganizada. Não há, a princípio, um formato que possa ser diretamente analisado por uma máquina sem uma pré-transformação adequada.

É nessa etapa que há o tratamento de dados ausentes, a remoção de dados duplicados e a tentativa de lidar com dados errados e inconsistentes.

2.1.1 Codificação, Normalização, Padronização

Primeiro deve-se procurar adaptar os dados para um formato em que os algoritmos de aprendizado de máquina possam ser utilizados. Itens qualitativos como, por exemplo, se o imóvel aceita pets ou não, se tem piscina ou não, entre outros, são abundantes no conjunto de dados e precisam ser ajustados. Para esses casos, por exemplo, usando a NBR como referência, há a sugestão de conversão para variáveis booleanas.

No entanto, nem todos os dados se resumem a “SIM” ou “NÃO” para serem convertidos diretamente no formato booleano, exigindo abordagens diferentes.Joshi (2016) menciona a possibilidade de usar o *Label Encoding*, que é uma forma de codificação que transforma variáveis categóricas em número. Por exemplo, a conservação de um imóvel pode estar representada como: ruim, regular, bom ou excelente. Com o *Label Encoding* é possível trazer esses dados para uma escala ordenada numérica, conforme indicado no exemplo da Figura 1.

Figura 1: Label Encoding aplicado

Sem Label Encoding		Ruim → 0	**Com Label Encoding**	
registro	Conservação do Imóvel	Regular → 1	registro	Conservação do Imóvel
0	Ruim	Bom → 2	0	0
1	Regular	Excelente → 3	1	1
2	Bom		2	2
3	Excelente		3	3
4	Regular		4	1

O grande problema é que nem sempre os dados são ordinais, por exemplo, às vezes só existe um campo String que descreve uma característica qualquer do imóvel como sauna, hidromassagem, piscina, campinho de futebol. Para esses casos, em que as características qualitativas não podem ser ordenadas, existe uma técnica de codificação mais adequada. Por exemplo, Muller e Guido (2016) usam o *One-Hot-Encoding*. A Figura 2 é uma demonstração de como o método pode ser aplicado para o caso de imóveis.

Figura 2: One Hot Encoding aplicado

Sem One-Hot-Encoding			**Com One-Hot-Encoding**			
registro	Característica		registro	tem_sauna	tem_piscina	tem_campo
0	Sauna		0	1	0	0
1	Hidromassagem		1	0	1	0
2	Piscina		2	0	0	1
3	Campinho de futebol		3	0	0	0

Além disso, muitos algoritmos são sensíveis à escala dos dados, tornando a padronização uma etapa importante no pré-processamento. Segundo DE AMORIM et al. (2023) alguns métodos para isso são o *Min-Max Scaling*, o *Standard Scaling* e o *Robust Scaling*.

No *Min-Max Scaling*, após a transformação, todos valores se situam numa escala de [-1,1], a fórmula aplicada para normalização é:

$$Valor\ Normalizado\ = \frac{Xi - Xmin}{Xmax - Xmin}$$

Onde,

- X_i é o valor atual
- X_{min} é o menor valor da amostra
- X_{max} é o maior valor da amostra.

Esse método tem como vantagem manter a real distribuição dos dados, porém tem como ponto negativo uma enorme sensibilidade a *outliers*. Já o *Standard Scaling*, segundo De Amorim et al. (2023) , é baseado na normalização *Z-score*. Ou seja, uma normal padrão, com média 0 e desvio padrão 1.

$$Valor\ Padronizado\ = \frac{X - \mu}{s}$$

Onde,

- X é o valor que se quer padronizar
- μ é a média amostral
- s é o desvio padrão amostral

Por último, segundo De Amorim et al. (2023) o *Robust Scaler* é um ótimo método para minimizar o efeito de *outliers*, já que ele se baseia em mediana e não na média. Dando menos peso aos valores mais discrepantes.

$$Valor\ Robust\ = \frac{Xi - Q2(X)}{Q3(X) - Q1(X)}$$

Onde,

- X_i é o valor que se deseja transformar
- Q1(X) é o valor delimitador do primeiro quartil (25% dos dados)
- Q2(X) é a mediana do conjunto de dados
- Q3(X) é o valor delimitador terceiro quartil (75% dos dados)

2.1.2 Redução de Dimensionalidade

É necessário, dentro do dataset, escolher quais atributos utilizar no meio de uma extensa lista de possibilidades. Géron (2019) menciona que o excesso de atributos pode deixar o treinamento muito lento, demandando mais poder computacional e, além disso, muitas características inúteis podem até atrapalhar a obtenção de uma boa solução. Uma alternativa oferecida pelo autor é o uso do PCA (*Principal Component Analysis*).

Verleysen (2005) menciona o conceito de *"Curse of Dimensionality"*, traduzindo, maldição da dimensionalidade, que, para ele, vem a impactar seriamente

a performance de um modelo. Quando há muitas dimensões, há dificuldades tanto de visualização quanto de processamento dos dados, já que os modelos se tornam excessivamente custosos, computacionalmente falando.

Géron (2019) menciona que o PCA é uma das técnicas de redução de dimensionalidade mais utilizadas atualmente. A técnica consiste em uma espécie de transformação ortogonal em que os atributos iniciais são transformados em um novo conjunto de atributos ortogonais, chamados de componentes principais. O Primeiro eixo, o PC1 é o eixo que é responsável pela maior quantidade de variância nos dados. O PC2 vem logo após, sendo o eixo responsável pela segunda maior quantidade de variância, seguindo a mesma lógica para todos os outros componentes principais. Com isso, em alguns casos, é possível que dezenas, ou, até mesmo centenas de atributos possam ser reduzidos a duas ou três dimensões que podem explicar, de forma satisfatória, boa parte da variância dos dados.

2.2 Remoção De Outliers

2.2.1 O problema do outlier no contexto imobiliário.

Lidar com *outliers* é o maior desafio na modelagem de preço de imóveis, principalmente em dados coletados via *scrapper*. Os motivos são inúmeros e serão descritos nesta presente seção.

A primeira grande dificuldade é que o preço ofertado pelo imóvel é diferente do preço efetivamente negociado. Segundo matéria do Estadão, feita por Damascena (2024), 61% dos negócios com imóveis no último trimestre de 2023 tiveram diferença entre o preço anunciado e o vendido. Além disso, uma matéria da revista Exame feita por Quesada (2024) revela que, em uma pesquisa do Datafolha, mais da metade dos proprietários afirmaram inflar os preços dos imóveis antes da venda e que apenas 16% definem o preço diretamente no nível que aceitariam fechar o negócio, sem abertura para negociações.

Um segundo desafio é a precificação. Em alguns casos, vendedores anunciam imóveis com valores muito acima do mercado, o que pode resultar em longos períodos de vacância, deixando o imóvel “encalhado”. Por outro lado, há situações em que, por necessidade de dinheiro imediato devido a emergências, os proprietários acabam oferecendo preços abaixo do valor de mercado, o que também compromete a análise.

Outro problema é o de acurácia dos dados. Como a maior parte dos dados são preenchidos por corretores imobiliários, há o risco de informações imprecisas. Por exemplo, um apartamento pode ser anunciado como tendo uma piscina que não existe, com informações incorretas sobre sua idade ou com o valor do condomínio desatualizado. Esse tipo de erro só poderia ser minimizado por meio de uma verificação manual individual de cada amostra, o que se torna inviável no contexto do uso de *scrappers*.

Além disso, o estado de conservação externa do imóvel é um aspecto fundamental, mas difícil de avaliar com base em dados isolados. Um prédio construído em 1970 pode ter passado por uma reforma recente e contar com uma boa gestão condominial, enquanto outro, com a mesma idade e características, pode estar em condições precárias. Assim, o estado de conservação exerce um peso significativo, que não pode ser adequadamente capturado apenas por meio da análise de dados brutos.

Por último, o estado de conservação interna do imóvel é também um fator crucial na determinação de seu valor. É possível que apartamentos com a mesma área, planta e andar, apresentem valores completamente distintos. Esse fenômeno pode ser atribuído, em primeiro lugar, à condição interna do imóvel. Um apartamento que passou por uma reforma recente, com acabamentos modernos e bem conservados, certamente terá um valor superior em comparação a outro que apresenta sinais de desgaste ou falta de manutenção, ainda que localizados em um mesmo edifício e possuindo as mesmas características.

2.2.2 Métodos matemáticos

A primeira abordagem para lidar com valores atípicos considerada neste trabalho foi o Isolation Forest, introduzido por Liu, Ting e Zhou (2008). Segundo os autores, em um espaço multidimensional, as anomalias, por serem diferentes e distantes dos padrões normais, são mais facilmente isoladas em comparação com dados normais. Para isso, os autores propõem uma metodologia baseada em árvores de decisão aleatórias.

A segunda abordagem foi mais simples, é baseada no intervalo interquartil (IQR) para a detecção de *outliers* introduzida por Tukey (1977). O método consiste em calcular o primeiro quartil, onde 25% das amostras residem, a mediana, que corresponde a 50%, e o terceiro quartil. Qualquer amostra que se encontre a uma

distância de 1,5 vezes a diferença interquartil (Q3 - Q1) será considerada como um alto risco de ser uma amostra fora do padrão, ou seja, um *outlier*. Portanto, o intervalo seguro é definido como [Q1−1.5×IQR,Q3+1.5×IQR].

Por último, segundo Wasilewska (2021), a regressão pode ser utilizada para suavizar e lidar com dados ruidosos. A ideia é encontrar pontos que estejam muito distantes da reta de valores previstos.

2.3 Seleção de Features

O PCA descrito na seção 2.1.2 pode ser utilizado para selecionar as características mais relevantes. Como explicado anteriormente, ao analisar as componentes principais é possível identificar e escolher as características com maior peso em cada componente.

Além disso, há o RFE (*Recursive Feature Elimination*), um método que se mostrou extremamente relevante neste presente estudo. Guyon et al. (2002) foi um dos pioneiros a descrever a lógica base de um algoritmo de recursão focado para seleção de características. Kuhn e Johnson (2013) descreveram os principais passos resumindo o modelo de Guyon et al. (2002) conforme elencado e traduzido na Figura 3.

Figura 3: Passos resumidos do RFE de Guyon et al. (2002).

1. Ajustar/treinar o modelo no conjunto de treinamento usando todos os P preditores.
2. Calcular o desempenho do modelo.
3. Calcular a importância das variáveis ou o ranking.
4. Para cada subconjunto de tamanho Si, onde i= 1....S faça:
    5. Mantenha as Si variáveis mais importantes
    6. [Opcional] Pré-processe os dados
    7. Tune/Treine o modelo na base de treinamento usando Si preditores
    8. Calcule a perfomance do modelo
    9. [Opcional] Recalcule os rankings para cada preditor
10. fim
11. Calcular o perfil de desempenho sobre os Si
12. Determinar o número apropriado de preditores (ex: o Si associado com a melhor perfomance)
13. Ajuste o modelo baseado no Si ótimo

(KUHN; JOHNSON, 2013, p. 494-495, tradução nossa).

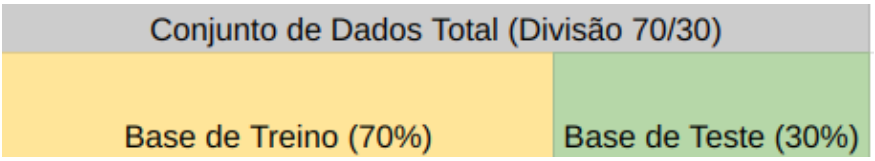
No geral, na implementação simplificada descrita por KUHN e JOHNSON (2013), o modelo é treinado de forma recursiva com todos os preditores, e sua performance é avaliada em cada etapa. A cada rodada, as variáveis mais importantes são selecionadas, enquanto as menos relevantes são removidas da amostra. A escolha da métrica a ser utilizada depende da implementação do modelo, mas podem ser empregadas diversas métricas como, por exemplo, o R^2, o RMSE ou a redução de impureza (gini ou entropia), para avaliar a contribuição de cada variável. Esse processo é iterativo, até que permaneça apenas o número previamente definido de variáveis.

2.4 Divisão da base de dados

Um dos maiores problemas ao tentar treinar um modelo de machine learning é o risco de *overfitting*. Isso é, quando o modelo criado possui um bom desempenho no conjunto de dados de testes, mas que não o mantém quando se depara com um conjunto de dados desconhecidos. Géron (2019) oferece uma solução para esse problema ao mencionar a técnica de *HoldOut*.

Essa técnica consiste em dividir a base de dados em duas: uma base de treino e uma de validação. Os percentuais de divisão variam de acordo com o interesse de quem vai modelar os dados, para esse trabalho será utilizado 70/30, que consistirá em 70% da base original para treinamento e 30% para validação como ilustrado na Figura 4.

Figura 4: Método HoldOut



Géron (2019) explica que a ideia do método *Holdout* é treinar as inúmeras possibilidades de modelo na base de treino, e testar com a base de validação, a modelagem que performar melhor na base de validação será a escolhida e posteriormente o modelo escolhido será ajustado considerando todo conjunto de dados.

2.5 Técnicas de Previsão de Preços

2.5.1 Análise de Regressão

Buscando obter um modelo de precificação se faz necessário investigar as relações entre diversos atributos que possam vir a influenciar o preço do imóvel. Características como o valor do condomínio, a área do imóvel, a localização entre outros fatores podem ser extremamente relevantes no preço final do imóvel. Essas variáveis serão chamadas de variáveis independentes e a relação entre essas características afetará o valor da variável dependente do estudo, o preço do imóvel.

> "A técnica mais utilizada quando se deseja estudar o comportamento de uma variável dependente em relação a outras que são responsáveis pela variabilidade observada nos preços é a análise de regressão." (NBR 14653-2, 2011)

Existem diversas formas de regressão, desde as lineares mais simples, até as mais complexas polinomiais. Para ilustrar, segundo Montgomery, Peck e Vining (2006), o modelo de regressão linear múltipla pode ser descrito como:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \ldots + \beta nxn + \varepsilon$$

Sendo,

y - variável dependente (preço do imóvel)

n – número de variáveis regressoras

β – coeficientes da regressão

x – variáveis independentes (ex: área, condomínio…)

ε – erro

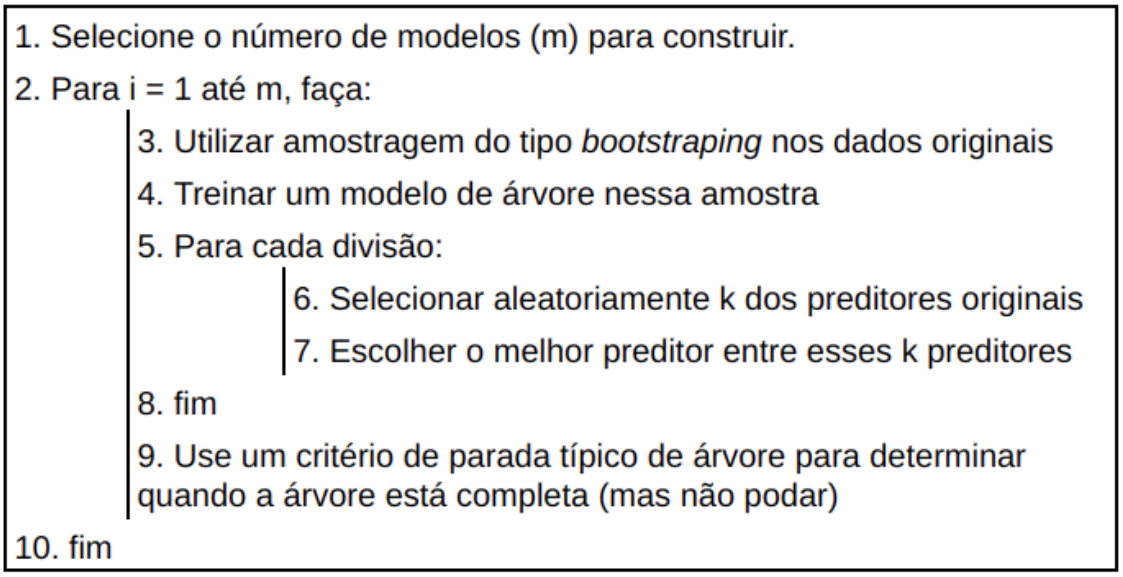
Além disso, a NBR 14653-2 (2011) afirma que, antes de aplicar uma regressão, é necessário checar alguns pressupostos como a normalidade do erro, homocedasticidade, não multicolinearidade, não-autocorrelação.

2.5.2. Árvore Aleatória (Random Forest)

A *random forest* é um tipo de estratégia que utiliza como base as árvores de decisão. No geral, as árvores são um modelo com baixo viés e alta variância, levando a um alto risco de *overfitting*.

Kuhn e Johnson (2013) mencionam que a estratégia da *random forest* visa minimizar isso ao combinar previsões de múltiplas árvores de decisão fracas, geralmente por meio de votação ou média para chegar na versão final do que está tentando se estimar. A estratégia é descrita pelos autores na Figura 5.

Figura 5: Algoritmo base de Random Forest



(KUHN; JOHNSON, 2013, p. 200, tradução nossa).

Bagging é uma técnica que utiliza-se de amostragem com reposição. Isso é especialmente útil, principalmente, nos casos em que não há uma alta disponibilidade de dados para montar múltiplas árvores. Kuhn e Johnson (2013) mencionam que além da amostragem com reposição, cada árvore montada varia as características selecionadas para predição, sendo distintas uma das outras.

A combinação da amostragem com reposição do *bagging* junto com variação de características gera árvores com um viés maior do que o de uma árvore de decisão simples, mas que, ao serem combinadas, resultam em um modelo de menor variância. Isso é útil para evitar o *overfitting*, já que um modelo de menor variância indica que o seu desempenho ao se deparar com dados desconhecidos não varia muito.

2.6 Avaliação do Modelo

Joshi (2016) menciona que existem diversas métricas para analisar a qualidade de um modelo de regressão. Dois exemplos que serão utilizados nesse trabalho são: O RMSE - *Root Mean Squared Error* (Erro Quadrático Médio da Raiz) e o *R-squared* (R-quadrado), também chamado de coeficiente de determinação.

O coeficiente de determinação R² é uma das principais métricas avaliadas para julgar o desempenho dos modelos de aprendizado de máquina. Charnet et al. (2008) descrevem o R² como “a proporção das variabilidades dos Y’s observados, explicada pelo modelo considerado”. Em outras palavras, essa métrica indica o quanto as características selecionadas de um imóvel, como, por exemplo, área e idade, podem vir a impactar no preço final.

$$R^2 = 1 - \frac{SQE}{SQT}$$

Sendo,

SQE - Soma dos Quadrados dos Erros

SQT – Soma dos Quadrados Totais

Porém, analisar somente o coeficiente de determinação R² pode ser enganoso. Ainda que o modelo venha a explicar uma parte da variação da variável, é possível que certos padrões de erro não sejam capturados corretamente. Montgomery, Peck & Vining (2006) dizem que, por exemplo, ao adicionar muitas variáveis no modelo, o coeficiente tende a aumentar, mesmo que a adição de variáveis não melhore o modelo. Por isso, além do R², é importante analisar com outras métricas como a raiz do erro quadrático médio (*Root Mean Squared Error*, RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Yi - \hat{Y}i)^2}{N}}$$

Sendo,

Yi - Valor da real da observação

Ŷi– Valor da previsão emitida pelo modelo

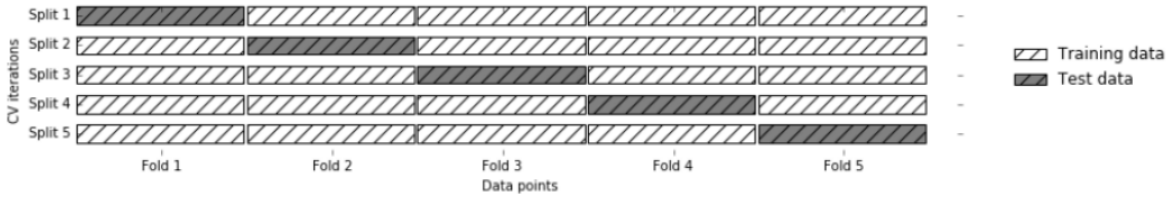
N = número total de observações

O RMSE será a principal métrica a ser avaliada neste presente trabalho, a busca é por uma minimização desse avaliador.

Uma forma eficiente de avaliar o desempenho do modelo em diferentes disposições dos dados, utilizando métricas como o RMSE mencionado acima, e ao mesmo tempo evitar o *overfitting*, é por meio da validação cruzada.

Muller e Guido (2016) afirmam que o método de validação cruzada mais utilizado é o *K-FOLD Cross Validation*. Seguindo os autores, o método consiste em dividir os dados em k subconjuntos, sendo o número k especificado pelo usuário. Nessa divisão, uma das k partições será reservada ao conjunto de teste, enquanto as demais ao conjunto de treino, como mostra a Figura 6. A ideia é executar o modelo em todas as k divisões e no final escolher o modelo que tiver em média o melhor desempenho nas mais variadas divisões. A busca tem como objetivo evitar ao máximo o *overfitting*.

Figura 6: K-Fold-Cross-Validation para k



(Fonte: Muller e Guido, p. 252, 2016)

3. Do Pré-Processamento à Análise: Etapas da Pesquisa Imobiliária

3.1 Estruturação

O projeto consistiu na análise de 3347 imóveis residenciais cadastrados em uma grande plataforma imobiliária brasileira que foram obtidos por uma ferramenta de *scraping*. Os dados foram coletados entre Maio de 2024 e Julho de 2024 de apartamentos localizados na cidade de São Paulo.

A análise se restringiu apenas a imóveis do tipo apartamento, devido a uma maior disponibilidade desses na plataforma imobiliária, além de também esta pesquisa partir do entendimento de que nem todos imóveis residenciais possuem as mesmas características. Portanto, apartamentos e casas demandam modelagem diferentes, justamente pelos fatores peculiares e inerentes de cada um, por isso a necessidade de separação e de um modelo individual distinto. Os bairros selecionados foram:

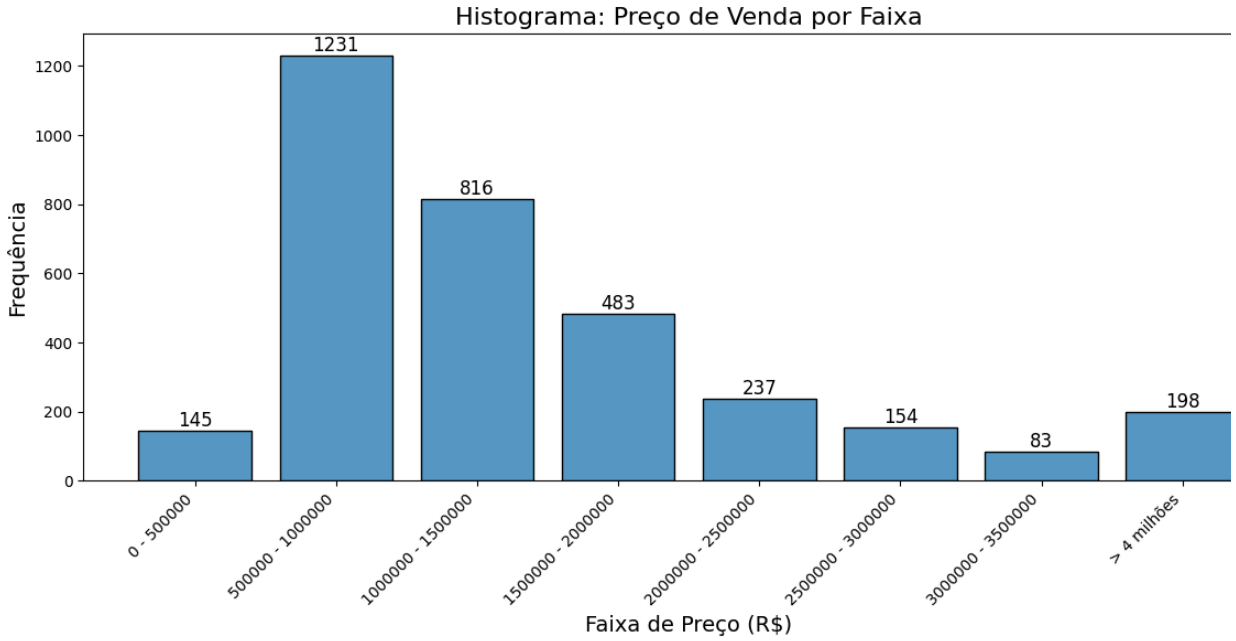
1. Vila Madalena
2. Vila Oímpia
3. Perdizes
4. Brooklin
5. Itaim Bibi
6. Jardim Paulista
7. Moema
8. Paraíso
9. Pinheiros
10. Vila Mariana
11. Vila Nova Conceição

São Paulo é uma megalópole com um número expressivo de bairros e sub-regiões. Houve a necessidade de redução do escopo, por isso se deu preferência a bairros com valores de metro quadrado mais elevados.

Os imóveis selecionados na amostra inicial, antes do pré-processamento, estão na faixa de R$ 280.000,00 (duzentos e oitenta mil reais) e R$ 18.850.000,00 (dezoito milhões, oitocentos e cinquenta mil reais), estando mais distribuídos na

faixa de R$ 500.000,00 (quinhentos mil reais) e R$ 1.000.000,00 (um milhão de reais) com cerca de 1231 amostras, conforme representados na Figura 7.

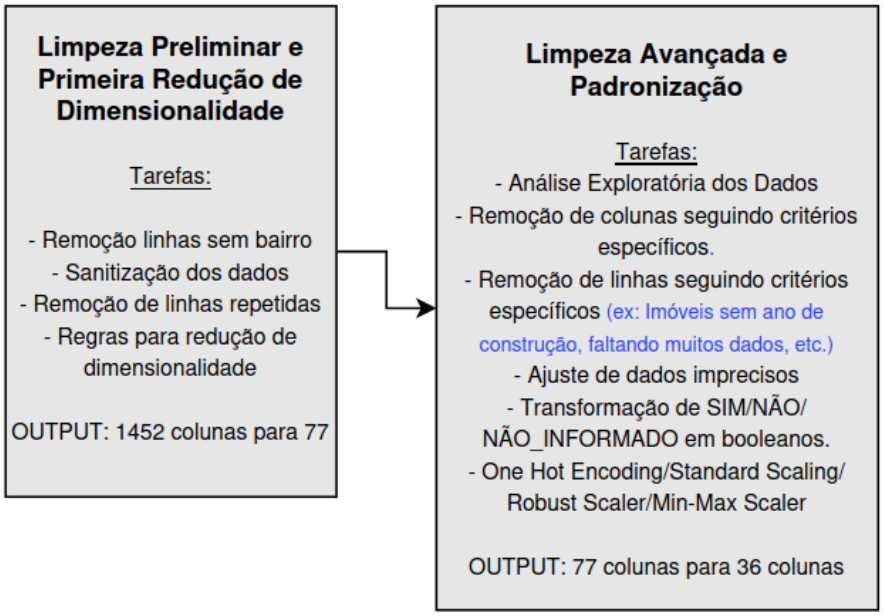
Figura 7 - Histograma Frequência de Preços



3.2 Pré-Processamento

A sequência de passos aplicados nesta etapa está resumida pela Figura 8 e será explicada com detalhes nesta presente seção.

Figura 8: Pré-Processamento Resumido



3.2.1 Limpeza Preliminar

Ação 1: Remoção de linhas sem bairro. Decidiu-se por remover todos os registros de imóveis que estavam sem bairro.

Ação 2: Sanitização dos Dados. Criação de uma função "limpar_bairro" com o objetivo de ajustar a formatação dos dados, essa função tem como objetivo :

- Remover todos os acentos
- Remover espaços em branco ao redor das palavras.

A explicação da lógica por trás dessa ação é que, após análise da base de dados, percebeu-se a existência de bairros registrados de forma despadronizada. Por exemplo, para o bairro Vila Olímpia, existiam registros das mais variadas formas como " Vila Olímpia", com espaço antecedendo a palavra, "Vila Olímpia", com acento, e também "Vila Olimpia", sem acento. Todos se referiam ao mesmo registro, por isso a necessidade de padronização.

Ação 3: Deduplicação. O *scrapper* em alguns momentos puxou o mesmo registro de forma duplicada. Para solucionar esse problema, criou-se uma função que verificava se existiam ID's do imóvel repetidos, caso sim, esses registros seriam removidos.

3.2.2 Redução de Dimensionalidade

O conjunto de dados inicial retirados do *scrapper* consistia em 1452 colunas, com uma alta dimensionalidade. Isso é problemático para modelos de aprendizado de máquina e muito provavelmente incorre no risco de "Maldição da Dimensionalidade", mencionado por Verleysen (2005).

Portanto, percebeu-se uma necessidade de reduzir o número de colunas para eliminar ruídos e informações desnecessárias. A ideia inicial foi usar alguma técnica de redução de dimensionalidade utilizando-se de métodos estatísticos, como o PCA, comentado na Seção 2.2 deste presente trabalho. Porém, Van Der Maaten (2009), cita que técnicas de redução de dimensionalidade clássicas, como o PCA, baseadas em métodos estatísticos também sofrem quando confrontadas com um número excessivo de dimensões.

O número de 1452 colunas vindos direto do *scrapper* é considerado elevado, então, antes de utilizar qualquer abordagem matemática, a primeira etapa do

pré-processamento foi uma tentativa de redução manual de dimensões, por isso, tentou-se localizar de forma visual, uma a uma, colunas contendo características irrelevantes ou redundantes que poderiam vir a ser removidas sem prejuízos.

3.2.2.1 Remoção Manual de Atributos - Regras Gerais

Durante a conferência, percebeu-se algumas possibilidades de remoção que foram ilustradas de forma exemplificativa na Figura 9:

- Colunas que continham informações de controle interno foram removidas (ex: photos/id ou photos/cover)
- Informações inseridas pelo usuário, sem padronização e que já tinham correspondente em outras partes do dataset também foram removidos (ex: photos/user/subtitle)
- Colunas que eram apenas links (ex: photos/url)

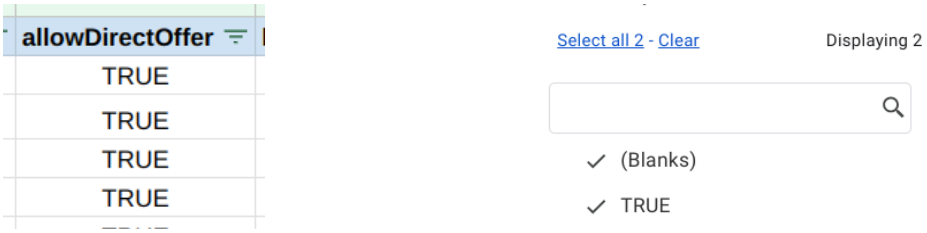
Figura 9: Exemplos de Remoções Manuais

photos/id	photos/user/subtitle	photos/url	photos/cover
53745841		894511-491.9034431797111363c7d3ce-db3c-407c-	FALSE
28458195		893730-109.83632641685482e08182248739b905e	FALSE
53719498	Sala	894408-170.72966634124475IMG9803.JPG	FALSE
53719820	Sala	894505-187.94628752194166MG0042.jpg	FALSE
53692381	Foto 17	894510-899.11918639407747564ad69-76b7-403f-9	FALSE
53766996	Sala	894509-716.2715342711531IMG0728.JPG	FALSE
53681247	Sala de Estar	894502-430.5422041018134IMG0003.JPG	FALSE

A demonstração exposta na Figura 9 é apenas um exemplo que se repete por todo conjunto de dados. Diversas colunas de controle interno foram removidas ao longo da coleção como, por exemplo, o código do corretor, o nome do corretor, a id de parceiros, entre outros.

Continuando a remoção, constatou-se a existência de colunas com todos parâmetros contendo um único valor booleano, ou seja, ou todos valores eram *TRUE* ou todos eram *FALSE*, no fim, não havia nenhuma agregação de informação, como ilustrado na Figura 10. Essas colunas também foram removidas.

Figura 10: Exemplos de Colunas com um único valor booleano



Dando seguimento a análise de características dos imóveis, percebeu-se a existência de muitas colunas com informações inválidas e algumas redundantes. Por exemplo, a Figura 11, mostra a existência de 2 colunas referentes a mesma característica (Varanda), isso se repete para diversos outros atributos, com muitos elementos duplicados desnecessariamente.

Além disso, certos atributos apresentaram valores booleanos inconsistentes, por exemplo, em certos casos, mesmo um apartamento possuindo varanda, a sua coluna mostrava o registro de "*False*". A constatação do equívoco foi feita após verificação manual, abrindo fotos do imóvel para avaliação e comparando o que estava sendo visto, com as informações registradas.

Abaixo a Figura 11 é uma exemplificação do que foi realizado e como ficou a coluna final. A ideia é remover as colunas repetidas e sem função (vermelho) e manter só as essenciais.

Figura 11: Etapas do Processamento

DADOS BRUTOS				APÓS PROCESSAMENTO 1		FINAL
Coluna 1	Coluna 1 - Existência	Coluna 1 - Repetida	Coluna 1 - Existência	Coluna 1	Coluna 1 - Existência	VARANDA
VARANDA	FALSE	Varanda	NAO_INFORMADO	VARANDA	NAO_INFORMADO	NAO_INFORMADO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	NAO_INFORMADO	VARANDA	NAO_INFORMADO	NAO_INFORMADO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	NAO	VARANDA	NAO	NAO
VARANDA	FALSE	Varanda	SIM	VARANDA	SIM	SIM
VARANDA	FALSE	Varanda	NAO_INFORMADO	VARANDA	NAO_INFORMADO	NAO_INFORMADO
VARANDA	FALSE	Varanda	SIM	VARANDA	SIM	SIM

O exemplo acima foi demonstrado com atributo "varanda", mas isso foi feito para todos os atributos existentes no dataset.

Após o término dessa etapa com a remoção de colunas de controle interno, remoção de colunas repetidas e inconsistentes, e outros processos aqui descritos; a quantidade de colunas foi reduzida de forma substancial, de 1452 para 77 itens.

3.2.2.2 Remoção Manual de Atributos - Regras Específicas

O corte realizado na etapa anterior foi expressivo. Porém, após análise minuciosa, constatou-se a possibilidade de uma redução ainda maior das 77 colunas resultantes. Em alguns atributos, após verificação manual, percebeu-se a existência de algumas características que eram mais livres, menos acuradas, pois eram preenchidas diretamente pelos corretores. Por isso, não havia uma garantia de exatidão na informação, existia uma margem maior para subjetividade. Por não haver uma forma objetiva de comprovar a veracidade das informações fornecidas, exceto por visitas presenciais, optou-se por eliminar esses atributos. A enumeração abaixo apresenta alguns exemplos de características removidas devido à sua natureza subjetiva ou de difícil verificação:

- Vista Livre
- Jardim
- Rua Silenciosa
- Sol da Manhã
- Sol da Tarde
- Janela Antiruído
- Luminosidade Natural
- Área Verde

A Figura 12 é um exemplo da importância dessa etapa, alguns corretores colocavam o atributo “Área Verde” como positivo, mas existia uma diferença gritante no que cada um considerava “Área Verde”. Para uns, bastava a presença de um simples vaso de planta e o atributo receberia o “*check”,* como na Figura 12a, para outros, de fato, existia a necessidade de um jardim robusto, conforme Figura 12b. Essa presença de subjetividade, junto com falta de padronização poderia vir a prejudicar o modelo.

Figura 12: Comparação Atributo: Área Verde

(a) Imóvel 1 (b) Imóvel 2

Ademais, também foi constatada existência de diversas colunas que eram voltadas apenas para imóveis de aluguel, não para venda, logo também foram removidas, conforme Figura 13.

Figura 13: Colunas com parâmetros que não servem para a análise de compra.

visitsEnabled	rentOnTermination	rentPrice
TRUE	FALSE	0
TRUE	FALSE	3100
TRUE	FALSE	0
TRUE	FALSE	3000
TRUE	FALSE	0
TRUE	FALSE	0

Após a aplicação de todas as regras elencadas nesta subseção houve uma redução de 77 para 38 colunas. Sendo as colunas remanescentes, descritas na Figura 14.

Figura 14: Características preliminares remanescentes.

Informações Preliminares dos Imóveis			
1	Área	23	Playground
2	Ano de Construção	24	Piscina
3	Bairro	25	Churrasqueira
4	Próximo ao Metrô	26	Quadra Esportiva
5	Vagas de Estacionamento	27	Academia
6	Aceita Animais	28	Salão de Festas
7	Tipo (Apartamento ou Studio/Kitnet)	29	Salão de Jogos
8	Andar	30	Sauna
9	É cobertura?	31	Lavanderia no Prédio
10	Mobiliado	32	Espaço Gourmet na Área Comum
11	Valor do Condomínio (R$)	33	Brinquedoteca
12	Quartos	34	Elevador
13	Suítes	35	Porteiro 24 Horas
14	Banheiros	36	Preço de Venda (R$)
15	Banheira de Hidromassagem		
16	Quarto de Serviço		
17	Banheiro de Serviço		
18	Quarto Extra Reversível		
19	Closet		
20	Cozinha Americana		
21	Varanda		
22	Varanda Gourmet		

3.2.3 Análise Exploratória Inicial e Limpeza Avançada

Nesta etapa deste presente trabalho, acontece uma junção das segundas e terceiras fases da metodologia CRISP-DM, um famoso *framework* com abordagens para orientar e sugerir passos que contribuem com a mineração de dados. A segunda fase da metodologia é o entendimento dos dados (WIRTH, 2000). Neste passo da mineração existe uma análise exploratória inicial dos dados para entender como está a disposição dos registros, se há registros faltantes, se há inconsistências, formular as primeiras hipóteses e *insights*. Na terceira etapa, a preparação dos dados, segundo Wirth (2000), é onde ocorre, de fato, a limpeza e a formatação dos dados. É nesse momento que se prepara todo o conjunto para a etapa seguinte, a de modelagem.

Após breve análise exploratória, percebeu-se a necessidade de mais ajustes no conjunto, além disso foram criadas algumas regras para lidar com dados faltantes ou inconsistentes.

Ajuste 1: remover linhas (registros) com mais de 12 campos "NAO_INFORMADO".

Figura 15: Regra 1 de remoção.

Imóvel	Bairro	Área	Banheiros	Quartos	Banheira_Hidro	Box	Varanda	Piscina	Closet
1	Perdizes	90	2	2	NAO	SIM	NAO	NAO	NAO
2	Perdizes	66	1	2	NAO	SIM	NAO	NAO	NAO
3	Perdizes	100	1	2	NAO	SIM	NAO	NAO	NAO
4	Perdizes	157	5	4	NAO_INFORMA	NAO_INFORMA	NAO_INFORMA	NAO_INFORMA	NAO_INFORM
6	Perdizes	125	3	3	SIM	NAO_INFORMA	NAO	NAO	NAO
7	Perdizes	80	1	1	NAO	SIM	NAO	NAO	NAO
8	Perdizes	65	1	1	NAO	SIM	NAO	NAO	NAO
9	Perdizes	111	4	3	NAO	NAO	SIM	NAO	NAO

Para cada imóvel, existem diversos atributos que representam suas características como, por exemplo, piscina, salão de jogos, quadra esportiva, banheira, etc. No entanto, por ser um campo de preenchimento de responsabilidade do corretor de imóveis, alguns apartamentos estão mal preenchidos, com diversos campos "NÃO INFORMADOS", ou seja, dados ausentes, como mostra a Figura 15. Registros com muitos dados faltantes podem afetar a previsão, portanto decidiu-se por eliminá-los.

Cerca de 124 linhas continham mais de 12 atributos não informados e foram removidas.

Ajuste 2: Remover as linhas em que o andar ou o ano de construção do edifício estão nulos ou vazios. Linhas removidas nessa etapa: 2336.

Ajuste 3: Codificação. *One Hot Encoding* na coluna “Tipo”, que tinha dois valores STRING: "Apartamento" e "Studio ou Kitnet", conforme a Figura 16.

Figura 16: Coluna em que foi aplicado o One-Hot-Encoding.

Imóvel	Bairro	Tipo	Área	Banheiros
100	Perdizes	Apartamento	80	2
101	Perdizes	Studio ou Kitnet	30	1
102	Perdizes	Apartamento	100	1
103	Perdizes	Apartamento	125	5
104	Perdizes	Apartamento	150	3
105	Perdizes	Studio ou Kitnet	27	1
106	Perdizes	Apartamento	65	1
107	Perdizes	Studio ou Kitnet	22	4

Além disso, implementou-se uma função chamada "codificar_colunas_ binarias", que tinha o objetivo de transformar valores {"SIM","NÃO"} em 1 e 0, como representado na Figura 17.

Figura 17: Antes e Depois da codificação.

					ANTES				
Imóvel	Bairro	Área	Banheiros	Quartos	Banheira_Hidro	Box	Varanda	Piscina	Closet
1	Perdizes	90	2	2	NAO	SIM	NAO	NAO	NAO
2	Perdizes	66	1	2	NAO	SIM	NAO	NAO	NAO
3	Perdizes	100	1	2	NAO	SIM	NAO	NAO	NAO
4	Perdizes	157	5	4	NAO_INFORMA	NAO_INFORMA	NAO_INFORMA	NAO_INFORMA	NAO_INF
5	Perdizes	125	3	3	SIM	NAO_INFORMA	NAO	NAO	NAO
6	Perdizes	80	1	1	NAO	SIM	NAO	NAO	NAO
7	Perdizes	65	1	1	NAO	SIM	NAO	NAO	NAO
8	Perdizes	111	4	3	NAO	NAO	SIM	NAO	NAO
					DEPOIS				
Imóvel	Bairro	Área	Banheiros	Quartos	Banheira_Hidro	Box	Varanda	Piscina	Closet
1	Perdizes	90	2	2	0	1	0	0	0
2	Perdizes	66	1	2	0	1	0	0	0
3	Perdizes	100	1	2	0	1	0	0	0
4	REGISTRO REMOVIDO ETAPA 1								
5	Perdizes	125	3	3	1	NAO_INFORMA	0	0	0
6	Perdizes	80	1	1	0	1	0	0	0
7	Perdizes	65	1	1	0	1	0	0	0
8	Perdizes	111	4	3	0	0	1	0	0

Ajuste 4: Remoção de colunas com muitos dados do tipo “NAO_INFORMADO”.

Na etapa seguinte, Ajuste 5, será feita uma conversão de "NAO_INFORMADO" para "NÃO", porém há um risco associado a essa ação que deve ser mitigado neste Ajuste 4. A ausência de informações sobre determinadas características no cadastro imobiliário nem sempre implica na inexistência desses atributos. Em alguns casos, essa falta de dados pode ser resultado de uma eventual displicência por parte do corretor no momento do preenchimento do cadastro.

Por isso, neste ajuste, visando preparar para a etapa posterior, optou-se por eliminar colunas que possuíam um alto grau de "NAO_INFORMADO". O limiar estabelecido foi que colunas possuindo acima de 20% de ausência deveriam ser removidas do presente estudo. Um exemplo das colunas que poderiam ser eliminadas está elencado na Figura 18.

Figura 18: Exemplo Ilustrativo do Ajuste 4.

Imóvel	Bairro	Área	Quartos	Banheira_Hidro	Varanda	Salao de Jogos	Cozinha Americana
200	Vila Mariana	80	3	0	0	0	NAO_INFORMADO
201	Vila Mariana	60	2	0	0	0	NAO_INFORMADO
202	Vila Mariana	76	2	NAO_INFORMA	0	NAO_INFORMA	NAO_INFORMADO
203	Vila Mariana	110	4	1	1	NAO_INFORMA	NAO_INFORMADO
204	Vila Mariana	150	4	1	1	1	0
205	Vila Mariana	28	1	0	1	NAO_INFORMA	0
206	Vila Mariana	52	2	0	1	NAO_INFORMA	1

Por esse critério, as colunas “Salão de Jogos” e “Cozinha Americana” foram removidas, por possuírem um percentual acima de 20% de ausência.

Ajuste 5: Converter atributo “NÃO_INFORMADO” para “NÃO” (bool 0).

Ajuste 6: Ajustes em inconsistências dos atributos. A Figura 19 abaixo mostra alguns exemplos de modificações realizadas.

Figura 19: Exemplo de Ajustes.

Coluna Problema	Descrição e Ajuste (Após Análise Exploratória)
Preço Condomínio	Existiam dados com condomínio preenchidos erroneamente, por exemplo, existia um condomínio com valor de R$ 690 mil. Critério de remoção: Se valor do condomínio < 55 ou > 16000, deletar registro. (OBS: foi checada a possibilidade do condomínio realmente ser abaixo de R$ 55, porém após análise manual, essa hipótese foi descartada)
Andar	Dois ajustes foram realizados: Remoção de andares acima de 80. Essa função foi criada, pois existiam registros com andares superiores a 80. O que após análise visual das fotos, se mostraram falsos. Remover andares abaixo ou iguais a 0. Cogitou-se a possibilidade do andar 0 ser do tipo “térreo”, mas após checagem dos links, essa hipótese foi descartada.
Quartos	Foi constatada a presença de registros com zero quartos. Após análise manual das amostras, percebeu-se que, na verdade, todos esses registros com “zero” eram studios. Para não perder os registros, converteu-se de 0 para 1.

Por fim, a próxima etapa é a de normalização dos dados, a fim de tentar deixar todos os registros em escalas similares, evitando distorções nas análises.

3.2.3 Normalização

Como a regressão utilizada em uma das etapas da modelagem é sensível à escala e distribuição, optou por realizar as seguintes transformações, conforme Figura 20.

Figura 20: Transformações Realizadas.

Característica	Método Utilizado
Banheiros	Min-Max Scaler
Quartos	Min-Max Scaler
Ano de Construção	Min-Max Scaler
Vagas de Estacionamento	Min-Max Scaler
Suítes	Min-Max Scaler
Valor do Condomínio (R$)	Standard Scaler
Área	Standard Scaler
Andar	Robust Scaler
Tipo	One-Hot-Encoding

Variáveis discretas, como banheiros, quartos, ano de construção e vagas de estacionamento, foram transformadas utilizando o *Min-Max Scaler*. No entanto, como esse método é sensível à escala, optou-se pelo *Standard Scaler* para variáveis contínuas com maior amplitude, como o valor do condomínio e área. Por fim, a variável andar foi escalonada com o *Robust Scaler*, que dá mais peso à mediana e é menos influenciado por *outliers*.

3.3 Pré-Modelagem

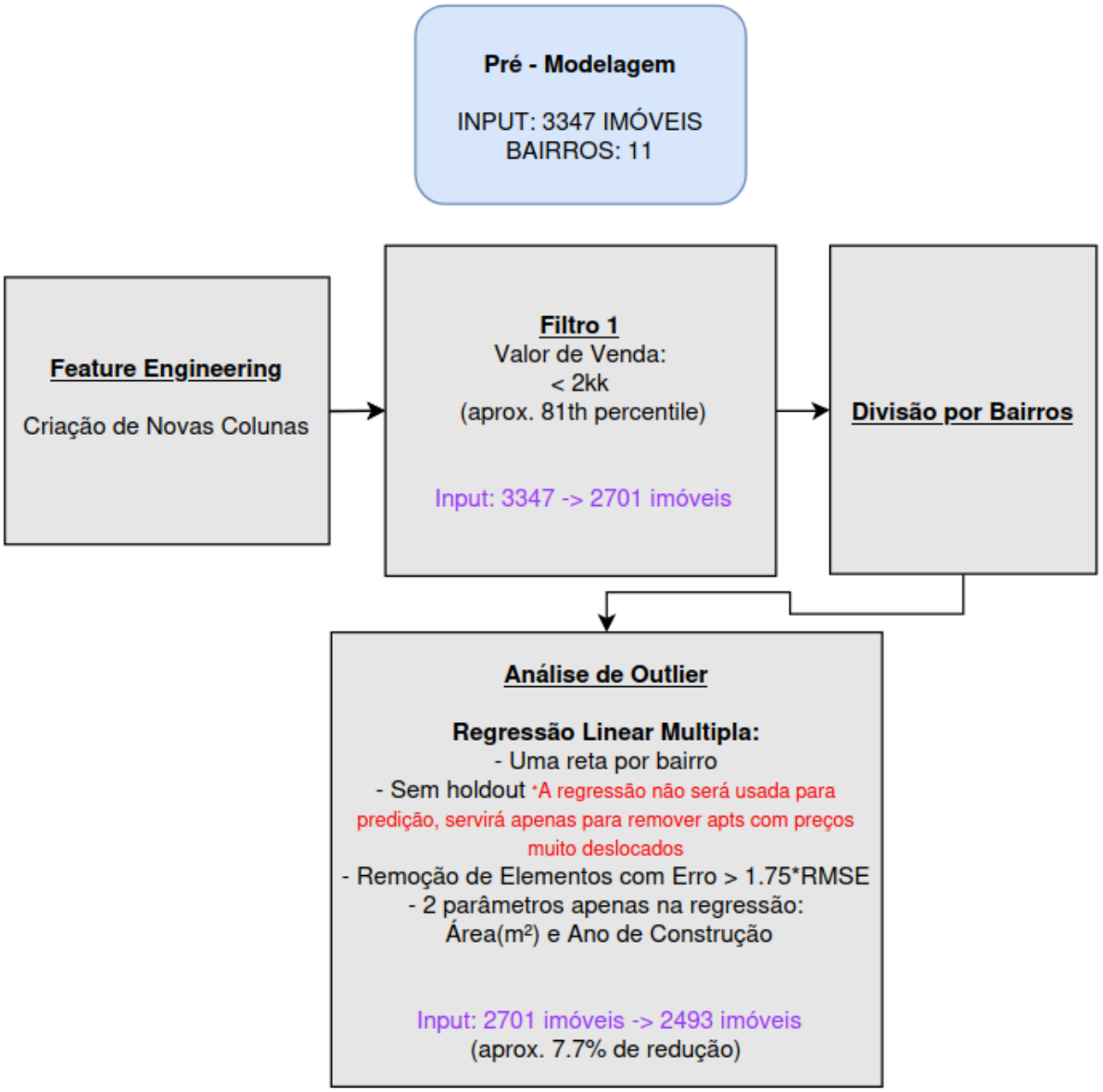
Segundo Wirth (2000), a etapa de preparação de dados na metodologia CRISP-DM engloba tanto a limpeza e a formatação dos dados, explicados na seção anterior, quanto a criação de novas características e remoção de *outliers*, a serem demonstradas nesta seção.

Já a etapa de modelagem é onde se adotam estratégias com os dados, como o uso de filtros, divisões, e também onde se selecionam as técnicas de aprendizado de máquina mais adequadas a serem utilizadas no contexto do projeto. É nessa etapa também que busca-se os parâmetros e hiperparâmetros mais adequados ao modelo.

Wirth (2000) menciona que existe uma proximidade muito grande entre preparação dos dados e a modelagem, segundo o autor é comum durante a etapa de modelagem perceber a necessidade de ajuste nos dados ou ter novas ideias para adaptações nos dados.

Como, na estratégia adotada neste trabalho, houve uma sobreposição entre alguns elementos da etapa de pré-processamento, como a remoção de *outliers*, junto com alguns ajustes e estratégias típicas de início da fase de modelagem, optou-se por criar uma etapa intermediária, apenas para fins de organização, denominada “pré-modelagem” e sintetizada pela Figura 21.

Figura 21: Pré - Modelagem

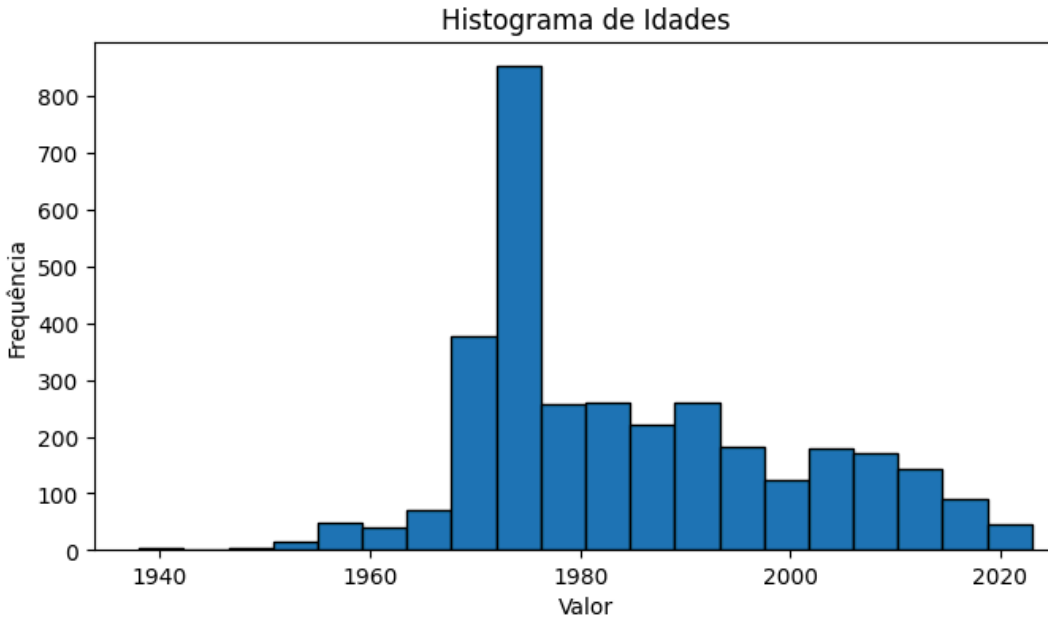


3.3.1 Feature Engineering - Criação de Novas Características

Nargesian et al. (2017) menciona a importância da criação de características para a melhora na performance preditiva de um modelo. Para os autores, a criação de características pode envolver aplicação de funções aritméticas ou de agregação nas características existentes para poder gerar novas. Essa manipulação pode ser baseada em expertise do analista, como também tentativa e erro.

As duas primeiras features criadas foram agrupamentos de idade. Uma análise exploratória inicial percebeu-se que há uma disposição muito variada de idade dos imóveis, conforme a Figura 22.

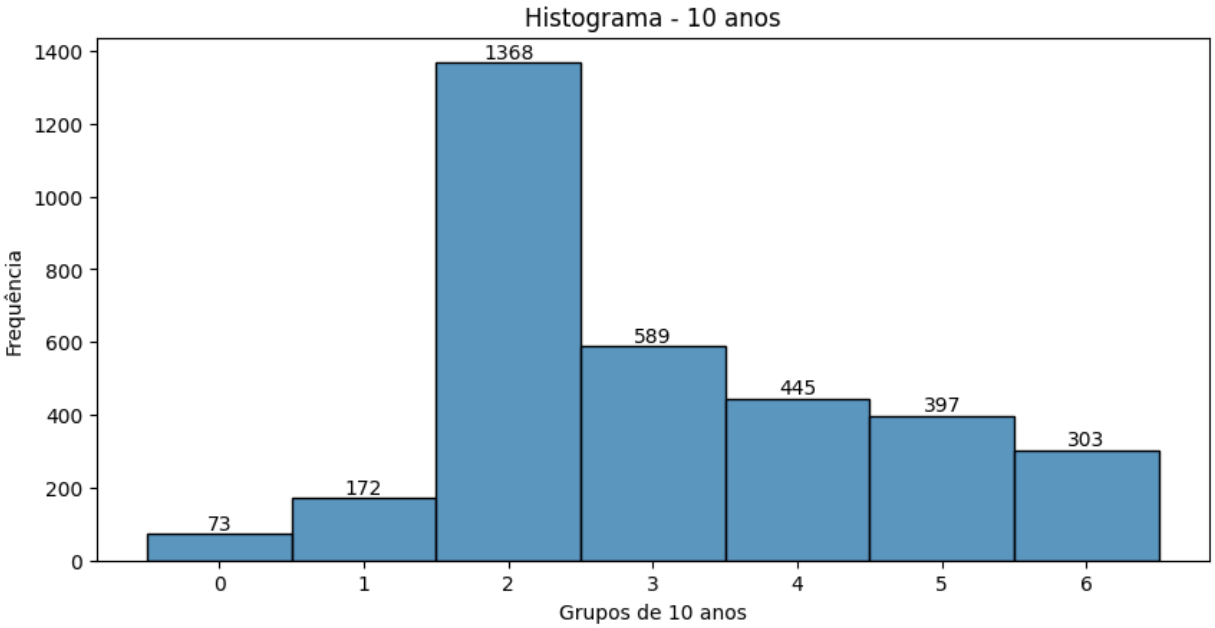
Figura 22: Histograma - Idade dos Imóveis.



É possível perceber uma concentração maior de imóveis relativamente antigos, com cerca de 40 a 50 anos. Isso é compreensível visto que São Paulo, por ser o centro econômico nacional, foi uma das primeiras cidades do país a ter uma grande expansão na construção de apartamentos residenciais.

Uma possível opção, na tentativa de gerar novas características, é a divisão em grupos de faixa de idade, o primeiro grupo criado contém intervalos de períodos de 10 anos e está representado conforme a Figura 23.

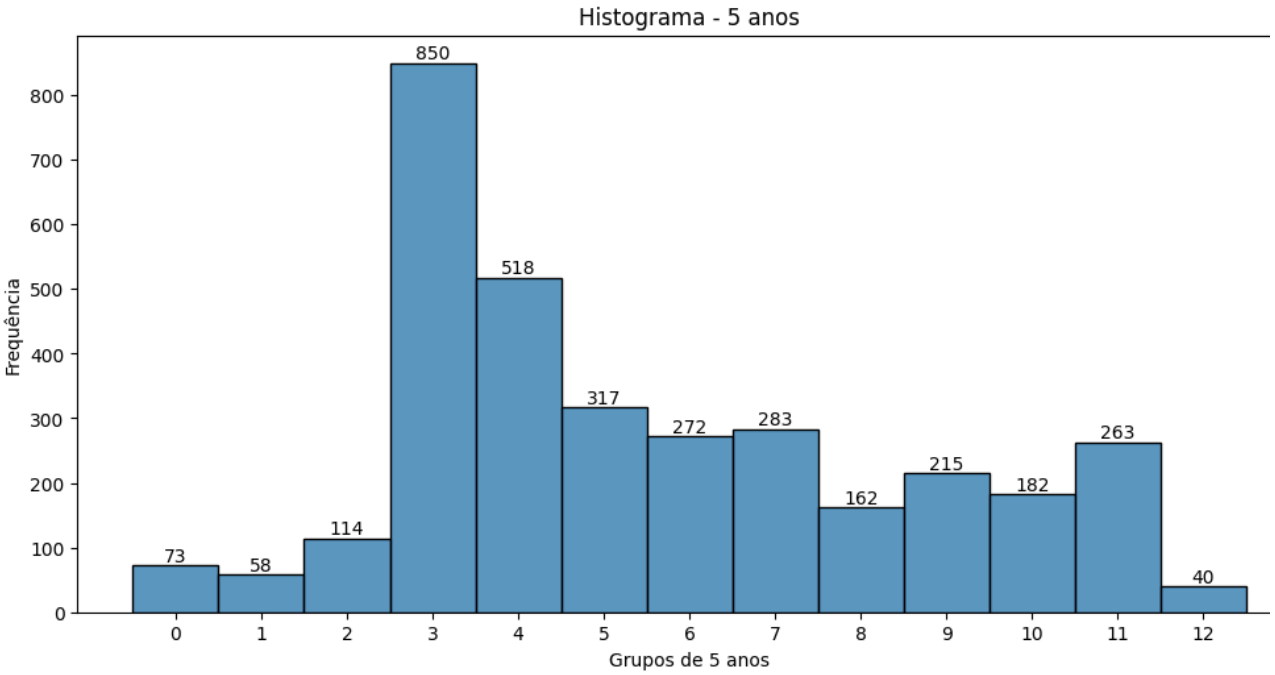
Figura 23: Grupamento de 10 anos



Agrupamento	Intervalo de Construção
0	< 1960
1	1960 ≤ construção < 1970
2	1970 ≤ construção < 1980
3	1980 ≤ construção < 1990
4	1990 ≤ construção < 2000
5	2000 ≤ construção < 2010
6	≥ 2010

Além desse agrupamento, criou-se também um agrupamento de faixas de idade intervaladas em cinco anos. Para este agrupamento, representado pela Figura 24, houve um pequeno ajuste. Embora a maioria dos intervalos sejam de cinco anos, algumas faixas como a primeira, penúltima e última, foram estabelecidas com intervalos distintos.

Figura 24: Grupamento de 5 anos



Agrupamento	Intervalo de Construção
0	< 1960
1	1960 <= ano < 1965
2	1965 <= ano < 1970
3	1970 <= ano < 1975
4	1975 <= ano < 1980
5	1980 <= ano < 1985
6	1985 <= ano < 1990
7	1990 <= ano < 1995
8	1995 <= ano < 2000
9	2000 <= ano < 2005
10	2005 <= ano < 2010
11	2010 <= ano < 2020
12	ano >= 2020

Uma outra opção de criação de características foi o agrupamento com base em especificidades do imóvel. Por exemplo, foram criadas novas variáveis, como amenidades internas e amenidades externas, que são descritas em mais detalhes na Figura 25.

Figura 25: Novas Características Agrupadas

Nova Coluna	Descrição
soma_amenidades_internas1	Banheira de Hidromassagem + Varanda Gourmet + Closet + Apartamento Cobertura + Varanda + Mobiliado
soma_amenidades_internas2	Banheira de Hidromassagem + Closet + Varanda Gourmet + Mobiliado
soma_amenidades_internas3	Varanda + Mobiliado + Closet
soma_amenidades_internas4	Banheira de Hidromassagem + Closet + Varanda Gourmet
soma_amenidades_externas1	Piscina + Academia + Salão de Festas + Sauna + Brinquedoteca
soma_amenidades_externas2	Piscina + Academia + Salão de Festas + Sauna
soma_amenidades_externas3	Quadra Esportiva + Brinquedoteca + Playground

Cada característica presente na residência foi codificada com o valor 1, indicando sua existência. A soma dessas características foi sintetizada numa nova coluna, contendo o número total de características internas e externas presentes no edifício analisado.

Por exemplo, se o apartamento tem varanda, é mobiliado, mas não tem closet, ele vai ter o valor de “soma_amenidades_internas3” igual a 2, pois tem duas características presentes.

Optou-se por mesclar diferentes arranjos de possibilidades, a lista final das características geradas pode ser visualizada na Figura 26. A etapa de "*feature engineering*" resultou na criação de 9 novas colunas, que variam desde o agrupamento por faixas etárias até a inclusão de diferentes atributos internos e externos.

Figura 26: Novas Características Agrupadas

Características Finais dos Imóveis			
1	Área	23	Piscina
2	Ano de Construção	24	Churrasqueira
3	Bairro	25	Quadra Esportiva
4	Próximo ao Metrô	26	Academia
5	Vagas de Estacionamento	27	Salão de Festas
6	Aceita Animais	28	Sauna
7	Tipo (Apartamento ou Studio/Kitnet)	29	Lavanderia no Prédio
8	Andar	30	Espaço Gourmet na Área Comum
9	É cobertura?	31	Brinquedoteca
10	Mobiliado	32	Elevador
11	Valor do Condomínio (R$)	33	Porteiro 24 Horas
12	Quartos	34	Preço de Venda (R$)
13	Suítes	35	Idade (Grupos de 5 anos)
14	Banheiros	36	Idade (Grupos de 10 anos)
15	Banheira de Hidromassagem	37	soma_amenidades_internas1
16	Quarto de Serviço	38	soma_amenidades_internas2
17	Banheiro de Serviço	39	soma_amenidades_internas3
18	Quarto Extra Reversível	40	soma_amenidades_internas4
19	Closet	41	soma_amenidades_externas1
20	Varanda	42	soma_amenidades_externas2
21	Varanda Gourmet	43	soma_amenidades_externas3
22	Playground		

	Características Originais
	Características Criadas

3.3.2 Filtragem e Agrupamento por Bairros

A primeira estratégia de filtragem utilizada foi eliminar imóveis acima de R$ 2 milhões (cerca do 80º percentil). A razão por trás dessa ação foi a percepção, após experimentações, de uma piora do modelo quando incluídos apartamentos de valores mais elevados.

O motivo é que por serem imóveis de padrão luxo, há peculiaridades específicas inerentes a esse tipo de empreendimento, difíceis de capturar, o que demandaria uma necessidade de mapeamento específico, fugindo do escopo deste presente estudo.

A segunda estratégia de filtragem parte do princípio de que cada bairro possui características próprias que agregam valor ao comprador, indo muito além do simples valor do metro quadrado da região. Por exemplo, para um bairro, a proximidade ao metrô pode ser um fator determinante, já para outro, esse fator pode não ter tanta relevância. Da mesma forma, certas características, como a presença de uma brinquedoteca, podem ser mais valorizadas em bairros voltados para famílias, enquanto em áreas mais orientadas para negócios essa característica pode vir a ter um menor impacto. Por isso, visando capturar essas peculiaridades optou-se por uma análise separada e a criação de um modelo específico para cada bairro da cidade de São Paulo.

3.3.3 Remoção de Outliers

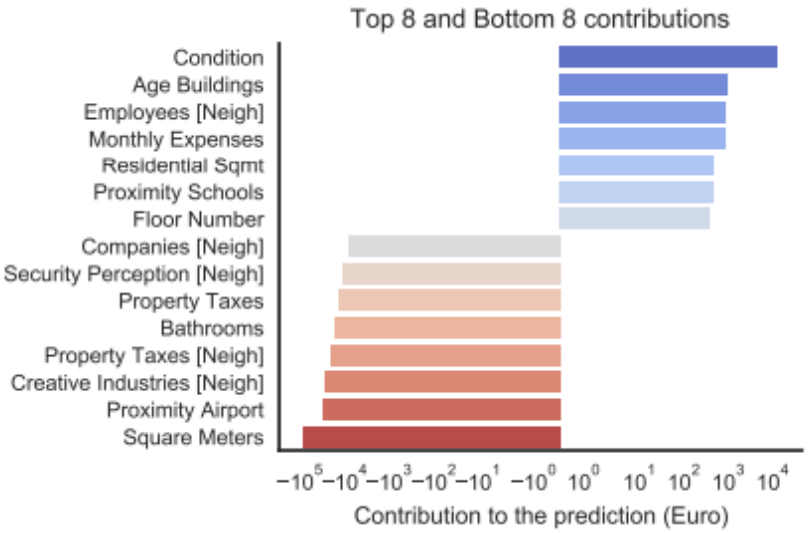
Como detalhado na Seção 2.2.1, lidar com *outliers* é um dos desafios mais significativos na modelagem deste trabalho. Como visto, esta situação pode ocorrer por várias razões, como discrepâncias entre o preço ofertado e o preço efetivamente vendido; imóveis anunciados com precificação incorreta; dados imprecisos fornecidos pelos corretores; ou fatores mais subjetivos que o modelo não consegue capturar, como a conservação externa ou interna do imóvel.

Neste trabalho, a primeira tentativa de eliminação de *outliers* foi realizada por meio do Isolation Forest, visando detectar amostras que divergiam do padrão. No entanto, o desempenho dessa abordagem não foi satisfatório, não havendo melhorias aparentes nos resultados. Como mais uma tentativa, aplicou-se a técnica de considerar como *outliers* os imóveis cujo preço estivesse acima ou abaixo de 1,5 vezes o intervalo interquartil (IQR), utilizando a fórmula $(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$. Essa abordagem, contudo, também não se mostrou eficaz.

A terceira estrateǵia foi a utilização de uma regressão linear para encontrar valores atípicos. Segundo Wasilewska (2021), a regressão pode ser uma ferramenta útil para suavizar e lidar com dados ruidosos. Para entender com quais características montar a regressão, um estudo conduzido por De Nadai e Lepri (2018) mostrou que, no geral, além da estado de conservação, as características

mais importantes de um imóvel são idade (correlação negativa) e metragem (correlação positiva), conforme ilustrado na Figura 27.

Figura 27 - Fatores que mais impactam no preço do imóvel



(DE NADAI; LEPRI, 2018)

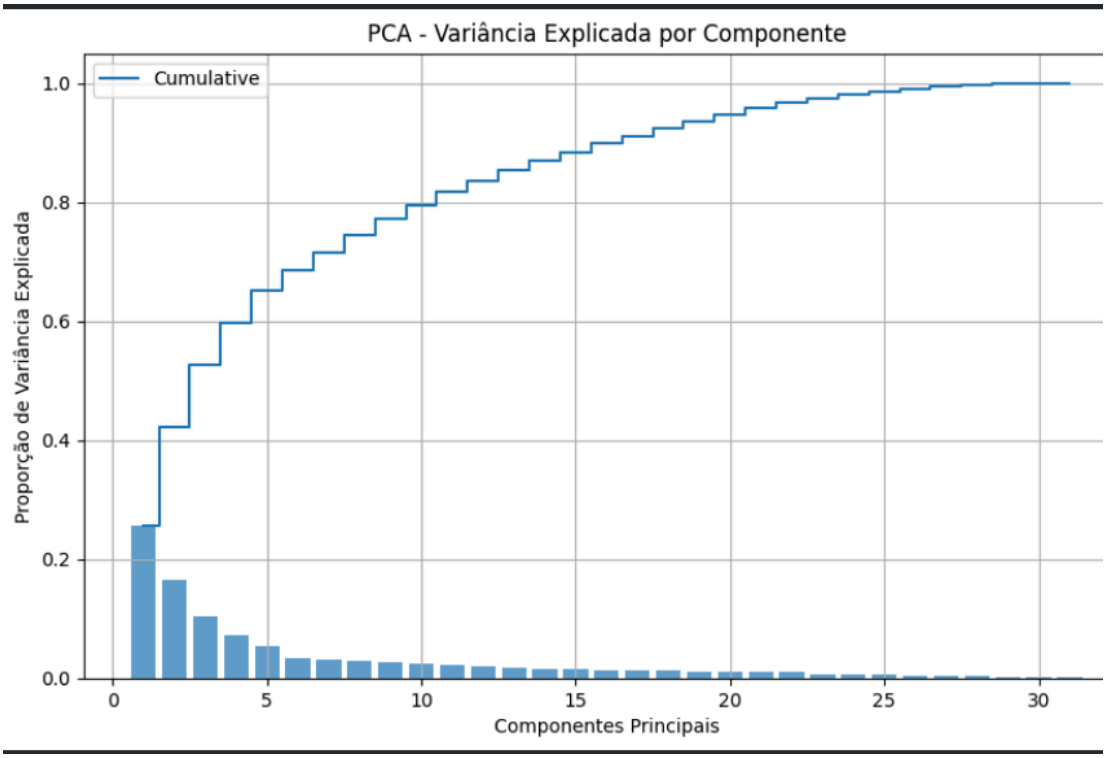
Decidiu-se criar um modelo subajustado (*underfitting*) que iria conter apenas essas duas variáveis independentes, idade e metragem. A ideia dessa regressão não seria a previsão de preços, mas sim, encontrar imóveis que estão diferindo muito do padrão em suas características “básicas”, eles seriam considerados valores atípicos.

Para essa análise, adotou-se um critério de remoção baseado no erro entre o valor previsto e o valor real, caso a diferença ultrapassasse 1,75 vezes o valor do RMSE (*Root Mean Square Error*) geral do modelo, a amostra correspondente seria excluída da análise. Esse valor de margem foi determinado baseado em testes experimentais que buscavam um limiar que eliminasse valores discrepantes, sem reduzir muitos dados do conjunto. A redução aproximada baseada nesse critério foi de apenas 7,7% das amostras.

3.3.4 PCA para redução de dimensionalidade

Após o pré-processamento, filtragem e remoção de *outliers* com as 34 características originais que podem vir a influenciar no preço do imóvel, se aplicou a técnica do PCA. Os resultados, elencados na Figura 28, mostram quanto foi o percentual de variância explicado por cada componente principal.

Figura 28 - Variância explicada por cada componente principal.



As seis primeiras componentes principais, explicaram cerca de 69% da variância total. Após isso, os ganhos não foram mais tão significativos. A Figura 29 abaixo mostra quais foram as duas características que mais impactaram cada componente principal.

Figura 29 - Variáveis que mais impactaram

Componente Principal (PC)	Variável 1	Peso 1	Variável 2	Peso 2
PC 1	Área	0.683989	Valor do Condomínio (R$)	0.649902
PC 2	Andar	0.833195	Ano de Construção	0.084447
PC 3	Piscina	-0.373699	Academia	-0.362409
PC 4	Banheiro de Serviço	-0.333834	Playground	-0.334069
PC 5	Quarto de Serviço	0.481818	Banheiro de Serviço	0.451284
PC 6	Próximo ao Metrô	0.557916	Aceita Animais	0.305155

O desempenho da Análise de Componentes Principais (PCA) não se mostrou muito eficiente para esse conjunto de dados. Por exemplo, para explicar 90% dos dados, precisaria-se usar 17 componentes principais, que é um número demasiadamente alto. Por isso, optou-se por usar o Recursive Feature Elimination (RFE), como será exposto posteriormente em mais detalhes na Seção 3.4.2 de modelagem com árvore.

3.4 Modelagem

3.4.1 Análise de Regressão

De início, a ideia era tentar implementar um modelo de regressão não só para remover *outliers*, mas também para a previsão dos preços em si. Alguns testes foram realizados e os resultados com a regressão não se mostraram satisfatórios. Após investigação para entender como melhorar os resultados, percebeu-se que alguns pressupostos da regressão estavam sendo violados, por isso decidiu-se não prosseguir com esse tipo de técnica de modelagem.

Conforme a NBR 14653-2 (2011) uma regressão linear exige alguns pressupostos como, por exemplo:

- Normalidade dos residuos
- Homocedasticidade
- Independência dos erros (não autocorrelação)
- Não multicolinearidade

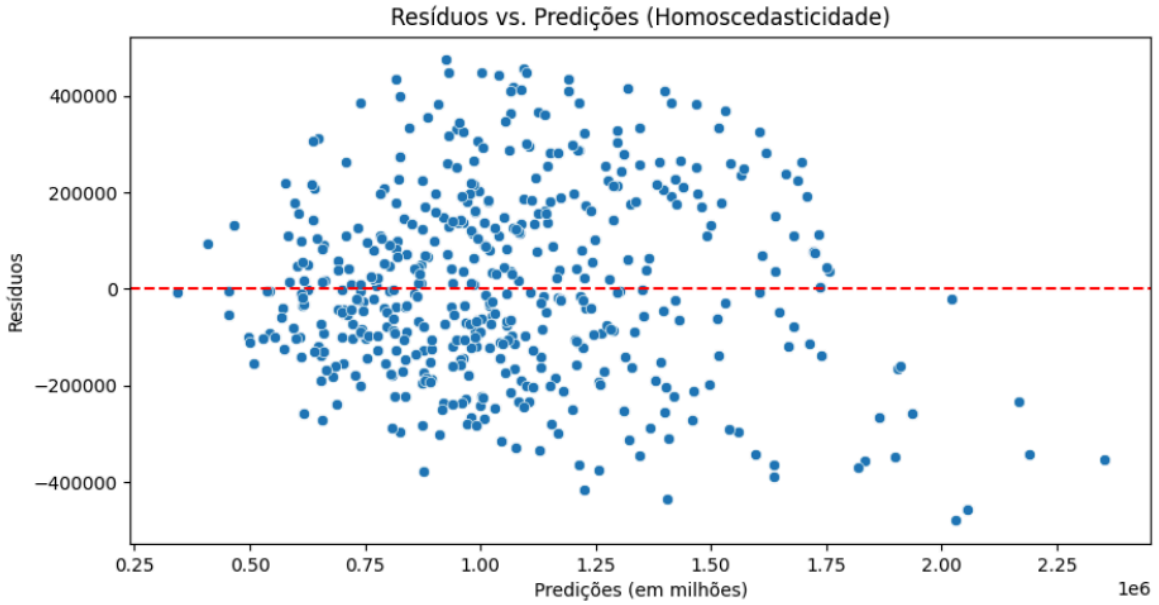
A multicolinearidade ocorre quando duas ou mais variáveis independentes estão correlacionadas entre si. Para evitar a multicolinariedade, no modelo de regressão, as características adicionadas na Seção 3.3.1, *feature engineering*, foram removidas. Ainda sim, nota-se que dois princípios foram claramente violados nos testes.

A primeira violação se diz respeito a premissa de homocedasticidade dos resíduos. Por esse pressuposto, os resíduos deveriam estar distribuídos de forma uniforme ao longo da amostragem para diferentes faixas de valores. A NBR 14653 (2011) menciona que uma opção para verificar a validade desse princípio é a elaboração de um gráfico “resíduos versus predições”.

“A análise gráfica dos resíduos versus valores ajustados, devem apresentar pontos dispostos aleatoriamente, sem nenhum padrão definido”. (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011).

O gráfico elaborado está disposto na Figura 30.

Figura 30 - Análise Gráfica Homocedasticidade



É possível notar que os erros não estão uniformemente distribuídos. Os imóveis na casa dos R$ 1 milhão possuem erros bem maiores do que os situados na faixa dos R$ 650 mil. Além disso, percebe-se que para imóveis de valor superior a R$ 1,75 milhão os erros são sempre negativos, havendo uma clara heterocedasticidade na disposição dos resíduos.

Constatou-se também que, mesmo com a remoção das características da Seção 3.3.1, ainda havia a existência de uma multicolinearidade muito grande entre as variáveis remanescentes e isso influencia negativamente na modelagem.

“Uma forte dependência linear entre duas ou mais variáveis independentes provoca degenerações no modelo e limita a sua utilização” (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011).

Montgomery, Peck & Vining (2006) citam que uma forma de verificar indícios de violação desse pressuposto é com o cálculo e constatação de valores altos do *Value Inflated Factor* (VIF).

Por exemplo, no conjunto de dados estudado, ao se calcular o VIF percebeu-se que a variável “Ano de Construção” tinha um valor extremamente alto, cerca de 26,17; conforme a Figura 31. Valores elevados implicam em um alto risco de dependência entre variáveis.

Figura 31 - VIF de algumas características

Característica	VIF
Ano de Construção	26.175202
Porteiro 24 Horas	10.613422
Vagas de Estacionamento	9.192663
Quartos	8.261835
Elevador	6.964488
Banheiros	6.069741
Área	4.437604
Salão de Festas	4.34195
Piscina	4.137212
Suítes	4.40956
Banheiro de Serviço	4.217003
Quarto de Serviço	3.920618
Aceita Animais	3.434722
Academia	3.659753
Próximo ao Metrô	3.01625
Playground	2.567066
Varanda	2.598813
Espaço Gourmet na Área C	2.032598
Valor do Condomínio (R$)	2.030235
Varanda Gourmet	2.019112

A dependência entre algumas dessas variáveis faz sentido, por exemplo, o valor elevado de “Ano de Construção” pode ser explicado uma vez que amenidades como piscina, academia, quadras esportivas, varanda gourmet e outras facilidades tendem a ser mais comuns em construções mais recentes e não tanto em imóveis mais antigos, logo há uma relação entre ano e outras variáveis.

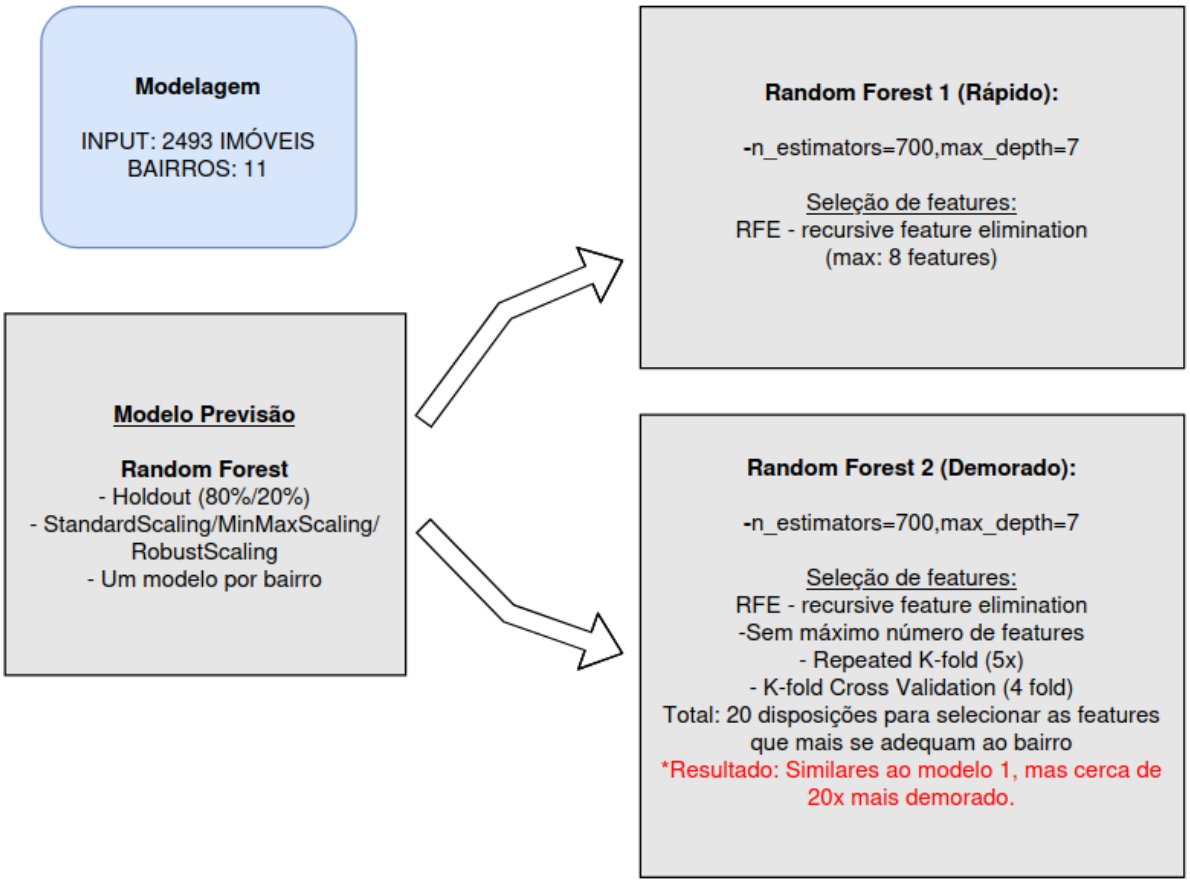
Vale ressaltar que a idade do imóvel não foi o único fator que chamou a atenção, outras variáveis também apresentam níveis alarmantes de dependência. Por exemplo, um apartamento com maior área(m²) tende também a ter mais banheiros, mais quartos, maior valor do condomínio, entre outros fatores que se apresentam elevados.

Por isso, diante desses padrões constatados nos dados, do risco de alta dependência entre variáveis, optou-se por não prosseguir com uma modelagem aprofundada de regressão, a ideia é priorizar algum modelo que não seja tão rígido em suas especificações e pressupostos.

3.4.2 Árvore de Decisão

Diante da descoberta da seção anterior, percebeu-se a necessidade de modelos mais flexíveis em seus pressupostos. Árvores de decisão se mostraram uma boa escolha, já que não exigem por padrão homocedasticidade, além de serem menos impactadas por *outliers* e multicolinearidade. A Figura 32 resume como foi feita a modelagem usando um modelo do tipo *Random Forest.*

Figura 32 - Modelagem



Preliminarmente, foram testados alguns modelos de árvore baseados em Boosting, como *Gradient Boosting, CatBoost* e *XGBoost*. Em seguida, avaliaram-se também modelos baseados em Bagging, como o *Random Forest*. Após algumas experimentações rápidas, o modelo que apresentou o melhor desempenho geral foi o *Random Forest* como será mostrado na Figura 33.

Figura 33 - Comparação de Modelos

Modelo	R^2	RMSE % (Relativo ao preço médio de venda)
Gradient Boosting	0.711	18.31%
XGBoost	0.722	18.19%
CatBoost	0.772	16.52%
Random Forest	0.792	15.78%

Entrando em mais detalhes em como se chegou a esse resultado, os passos seguiram o planejamento definido na etapa de Pré-Modelagem. Inicialmente, foram filtrados os imóveis com valor abaixo de R$ 2 milhões. Em seguida, os dados foram divididos por bairro, com a remoção de *outliers* em cada localidade. Posteriormente, foi desenvolvido um modelo de *Random Forest* específico para cada bairro, a fim de capturar as particularidades de cada área.

Após testar diferentes combinações de parâmetros, optou-se por duas abordagens para estruturar o *Random Forest*: uma mais simples e rápida e a outra mais criteriosa, com mais etapas, demandando mais tempo e poder computacional para sua execução, conforme a Figura 32.

A primeira abordagem, mais ágil, consistiu na divisão tradicional entre treino e teste para cada bairro, seguida da aplicação do método RFE (*Recursive Feature Elimination*) para selecionar as 8 variáveis mais relevantes em cada caso. Os resultados consolidados para todos os bairros estão apresentados na Figura 34, obtidos após a execução de aproximadamente 6 *seeds* diferentes.

Figura 34 - Modelo rápido de Random Forest

Desempenho Geral	
R^2	0.792
Preço Médio de Venda	R$ 1.030.813,80
RMSE	R$ 162.705,58
RMSE Relativo (ao Preço Médio)	15,78%
Amostras Treino/Teste	1990/503

A segunda abordagem, mais lenta, consistiu na divisão tradicional entre treino e teste para cada bairro, seguida da aplicação do método RFE *(Recursive Feature Elimination)* sem um número fixo de variáveis mínimas, usando o método de "*Repeated K-Fold*", ou seja, uma validação cruzada K-FOLD com 4 folds diferentes e 3 repetições, isso permitiu o modelo testar 12 disposições diferentes dos dados para, assim, selecionar as melhores features.

Os resultados que estão elencados na Figura 35, também foram obtidos após a execução de aproximadamente 6 *seeds* diferentes.

Figura 35 - Modelo lento de Random Forest

Desempenho Geral	
R^2	0.787
Preço Médio de Venda	R$ 1.030.813,80
RMSE	R$ 164.458,98
RMSE Relativo (ao Preço Médio)	15,95%
Amostras Treino/Teste	1990/503

Além do tempo de execução ser significativamente maior, cerca de 12 vezes mais, o modelo demonstrou pior capacidade de generalização com dados nunca antes vistos. Diante disso, optou-se por adotar a primeira abordagem, considerando seu melhor desempenho, simplicidade e também uma maior eficiência computacional. Os resultados da execução do modelo 1 esmiuçados para cada bairro estão dispostos na Figura 36.

Figura 36 - Resultados por Bairro - Média após 6 *seeds* diferentes

Bairro	R^2	Preço Médio de Venda (R$)	RMSE	RMSE relativo ao Preço Médio	Total Registros
Vila Madalena	0.666	1.108.179,49	183.654,94	16.55%	61
Vila Olímpia	0.723	929.153,33	143.778,36	15.50%	274
Perdizes	0.840	1.060.241,24	155.596,51	14.69%	450
Brooklin	0.901	1.056.792,61	139.863,54	13.25%	111
Itaim Bibi	0.651	1.173.718,18	217.182,78	18.50%	216
Jardim Paulista	0.874	1.089.465,23	150.678,21	13.86%	362
Moema	0.770	1.011.202,16	178.825,23	17.68%	135
Paraíso	0.787	1.045.691,08	179.212,21	17.16%	208
Pinheiros	0.758	1.046.021,23	174.436,27	16.69%	262
Vila Mariana	0.875	820.252,54	118.589,55	14.45%	260
Vila Nova Conceição	0.710	1.057.204,29	193.629,54	18.36%	154
Desempenho Geral	***0.792***	***1.030.813,81***	***162.705,59***	***15.79%***	***2493***

**Desempenho calculado considerando média ponderada

As oito características mais importantes por bairro apontadas pelo modelo de RFE foram descritas na Figura 37.

Figura 37 - Características mais importantes por bairro

Bairro	Características mais Importantes (Selecionadas pelo RFE)
Vila Madalena	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Suítes, Ano de Construção (Grupos de 10 anos), Ano de Construção (Grupos de 5 anos)
Vila Olímpia	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Ano de Construção (Grupos de 5 anos), soma_amenidades_internas1, soma_amenidades_externas1
Perdizes	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Suítes, Ano de Construção (Grupos de 5 anos), soma_amenidades_externas2
Brooklin	Área, Valor do Condomínio (R$), Ano de Construção, Vagas de Estacionamento, Ano de Construção (Grupos de 5 anos), soma_amenidades_externas1, soma_amenidades_externas2, soma_amenidades_externas3
Itaim Bibi	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, soma_amenidades_internas2, soma_amenidades_externas1, soma_amenidades_externas2

Figura 37 - Características mais importantes por bairro

Bairro	Características mais Importantes (Selecionadas pelo RFE)
Jardim Paulista	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Suítes, Ano de Construção (Grupos de 5 anos), soma_amenidades_externas1
Moema	Área, Valor do Condomínio (R$), Ano de Construção, Andar, Piscina, Vagas de Estacionamento, soma_amenidades_internas1, soma_amenidades_internas2
Paraíso	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Salão de Festas, Suítes, soma_amenidades_externas1
Pinheiros	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Vagas de Estacionamento, Ano de Construção (Grupos de 5 anos), soma_amenidades_externas1
Vila Mariana	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Ano de Construção (Grupos de 5 anos), soma_amenidades_externas1, soma_amenidades_externas3
Vila Nova Conceição	Área, Banheiros, Valor do Condomínio (R$), Ano de Construção, Andar, Piscina, Vagas de Estacionamento, soma_amenidades_externas1

De modo geral, o modelo apresentou um desempenho satisfatório, especialmente considerando o contexto imobiliário e os desafios discutidos no Item 2.2.1. O modelo obtido tem um coeficiente de determinação (R^2) de 0.792, ou seja, explica aproximadamente 79,2% da variação percebida nos preços dos imóveis. Para comparação, em um estudo semelhante sobre imóveis nacionais também obtidos via *web scraping*, Balczareki (2024) alcançou um R^2 de 0,70.

Por último, o RMSE relativo ao preço médio obtido neste presente trabalho foi de 15,79% e revela que o erro se manteve em níveis aceitáveis, especialmente para um modelo que não considera descontos aplicados em negociações, o estado de conservação dos imóveis, e outras subjetividades. Em alguns bairros, como Brooklin, Perdizes, Jardim Paulista e Vila Mariana, o modelo conseguiu estimar ainda melhor o valor dos apartamentos, apresentando valores de RMSE relativos mais baixos, situados entre 13% e 14% do valor total do imóvel.

4. Conclusão

Um dos grandes objetivos deste trabalho era buscar meios para tornar as decisões de compra de um imóvel em um ato mais racional. A busca era por implementar um modelo de boa performance na predição de preços, mas que acima de tudo fosse fácil de implementar, e, de preferência, utilizasse alguma técnica de ampla disponibilidade na literatura.

Precificar um imóvel com base apenas em dados brutos de uma grande plataforma imobiliária online apresenta inúmeros desafios, sendo o principal lidar com a discrepância entre o preço anunciado e o preço de venda, devido a fatores que os dados objetivos do modelo não conseguem capturar. As razões para isso são várias: a margem para negociação, o estado de conservação externa e interna do imóvel, a precificação incorreta — como imóveis que estão baratos demais para serem vendidos rapidamente devido a alguma urgência ou imóveis excessivamente caros que ficam anos sem vender — e dados preenchidos de forma inadequada na plataforma imobiliária.

A análise de regressão não se mostrou muito útil na tentativa de estimar o preço do imóvel, principalmente, devido à multicolinearidade das características e a heterocedasticidade dos resíduos, porém se mostrou uma ferramenta extremamente útil para remoção de imóveis com o preço fora do padrão, ou seja, *outliers*. Para precificação percebeu-se a necessidade de modelos mais flexíveis em seus pressupostos.

Árvores de decisão se mostraram uma boa escolha. Um modelo do tipo *random forest* foi criado para cada bairro, obtendo-se um desempenho médio geral satisfatório. O coeficiente de determinação (R^2) obtido foi de 0,792, ou seja, 79,2% da variação dos preços conseguiu ser capturada pelo modelo, e está com um desempenho satisfatório comparado a outros no mesmo contexto de *web scraping*, como o de Balczareki (2024), que alcançou um R^2 de 0,70.

O RMSE obtido foi de R$ 162.705,59 o que representa um RMSE relativo de 15,79% em relação ao preço médio dos imóveis analisados de R$ 1.030.813,81. O erro de predição está dentro dos níveis esperados. Por exemplo, ao se considerar o desconto médio das negociações, a famosa "pechincha", apenas ela é estimada em cerca de 11% do valor do imóvel (DAMASCENA, 2024). No entanto, além disso, há outros fatores mencionados anteriormente que influenciam na distorção do preço e

não são captados pelo modelo, como o estado de conservação externa e interna. Por isso, um RMSE de 15,79% pode ser considerado satisfatório.

Dada a simplicidade do modelo, e os diversos fatores que afetam a variação natural dos preços dos imóveis, os resultados obtidos indicam que, considerando as ressalvas expostas no item abaixo, o modelo pode ser um importante aliado na estimativa de preços dos imóveis. Apesar de toda modelagem ter sido feita baseada na cidade de São Paulo, os passos utilizados são genéricos e podem ser replicados com o devido cuidado e ajustes para diferentes localidades.

4.1 Limitações e Sugestões para próximos trabalhos

A primeira grande limitação é que a pesquisa está baseada apenas em alguns bairros de classe média alta da cidade de São Paulo, não sendo possível generalizar seus resultados para outras regiões do país. Além disso, os valores dos apartamentos foram restritos e estão limitados em até R$ 2 milhões devido ao fato que imóveis mais caros possuírem características próprias, mais subjetivas que necessitam de um modelo específico direcionados ao setor de alto padrão. O valor de R$ 2 milhões corresponde ao percentil 80 do conjunto total de dados extraídos.

Além disso, o modelo se restringe a analisar dados brutos. Há uma falha em análises que capturem características mais subjetivas, como se o imóvel passou por reformas ou qual é o estado interno do apartamento, entre outras nuances. Segundo De Nadai e Lepri (2018), a condição de um imóvel é um dos fatores que mais influencia em seu preço. Por isso, deve-se criar meios para tentar capturar essas características de alguma forma. Uma opção é o uso de algoritmos de aprendizado de máquina, como redes neurais convolucionais -CNNs-, para processar imagens dos imóveis, buscando assim meios de avaliar a condição de um imóvel de forma automática. Uma outra hipótese é uma checagem manual, imóvel a imóvel, atribuindo notas de conservação com base em critérios objetivos determinados previamente.

Vale ressaltar, que os resultados do modelo são baseados apenas em apartamentos e levando em consideração as características atuais do mercado, porém elas não são imutáveis ao longo do tempo, sendo necessário sempre a atualização e incorporação de novos fatores. Por exemplo, varanda gourmet era um item que não estava sendo levado em consideração nos imóveis mais antigos e

estão presentes nos imóveis novos, por isso, as modelagens devem ser constantemente refeitas, para capturar mudanças nos padrões de consumo.

Por último, uma grande limitação do modelo é que todas as árvores de decisão foram construídas levando em consideração o mesmo conjunto de características para todos os bairros. A performance do modelo poderia se beneficiar com a adição de características próprias, peculiares a cada bairro. Por exemplo, em certos bairros, pode-se assumir que a distância para a Avenida Paulista ou a Faria Lima seja algo relevante, em outros bairros a distância para grandes universidades, como a Universidade de São Paulo (USP), pode ter uma relevância no seu valor final. Além disso, a distância para parques como o Ibirapuera ou estações de metrô podem ser levadas em consideração para uma melhor modelagem em certos imóveis. Por isso, uma análise mais detalhada, específica, que considere as peculiaridades de cada bairro, incorporando características únicas relevantes, provavelmente, deve refletir em previsões de preço mais acuradas.

Referências Bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **NBR 14653-2: Avaliação de bens – Parte 2: Imóveis urbanos.** 2. ed. Rio de Janeiro, 2011.

BALCZAREKI, Yuri Potrich. **Precificação de imóveis em Florianópolis utilizando técnicas de aprendizado de máquina.** 2024. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) – Universidade Federal de Santa Catarina, Centro Tecnológico, Florianópolis, 2024. Disponível em: https://repositorio.ufsc.br/handle/123456789/255809. Acesso em: 11 nov. 2024.

CBIC - Câmera Brasileira da Indústria da Construção. **Indicadores Imobiliários Nacionais do 4º Trimestre de 2023.** 2024. Disponível em: http://www.cbicdados.com.br/menu/home/indicadores-imobiliarios-nacionais-4o-trimestre-de-2023. Acesso em: 07 de abril de 2024.

CHARNET, Roberto Carlos; FREIRE, Cláudio Antônio Lima; CHARNET, Eliane Maria Ribeiro; BONVINO, Hugo. **Análise de modelos de regressão linear – com aplicações**. 2. ed. Campinas: Editora da Unicamp, 2008. 356 p.

DAMASCENA, Breno. **Desconto médio no valor de venda de um imóvel é de 11%, indica estudo.** O Estado de S. Paulo, São Paulo, 19 fev. 2024. Disponível em: https://imoveis.estadao.com.br/compra/desconto-medio-no-valor-de-venda-de-um-imovel-e-de-11-indica-estudo/. Acesso em: 23 out. 2024.

DE AMORIM, Lucas B.V.; CAVALCANTI, George D.C.; CRUZ, Rafael M.O. **The choice of scaling technique matters for classification performance.** *Applied Soft Computing*, v. 133, p. 109924, 2023. DOI: https://doi.org/10.1016/j.asoc.2022.109924. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494622009735. Acesso em: 25 out. 2024.

DE NADAI, Marco; LEPRI, B. **The economic value of neighborhoods: predicting real estate prices from the urban environment.** In: *Anais da 2018 IEEE 5th*

International Conference on Data Science and Advanced Analytics (DSAA). 2018. p. 323-330.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2. ed. Sebastopol: O'Reilly Media, 2019. ISBN 1492032646; 9781492032649.

GUYON, Isabelle; WESTON, Jason; BARNHILL, Sarah; VAPNIK, Vladimir. **Gene selection for cancer classification using support vector machines.** Machine Learning, v. 46, n. 1, p. 389–422, jan. 2002. DOI: https://doi.org/10.1023/A:1012487302797.

JOSHI, Prateek. **Python Machine Learning Cookbook: 100 Recipes That Teach You How to Perform Various Machine Learning Tasks in the Real World**. Birmingham: Packt Publishing, 2016.

KUHN, Max; JOHNSON, Kjell. **Applied predictive modeling.** Nova York: Springer, 2013. p. 494-495. ISBN 978-1-4614-6848-6.

LIU, Fei Ting; TING, Kin Man; ZHOU, Zhi-Hua. **Isolation forest.** In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM). IEEE, 2008. p. 413–422. DOI: 10.1109/ICDM.2008.17.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to linear regression analysis.** 4. ed. Hoboken: John Wiley & Sons, Inc., 2006. 607 p.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning with Python: A Guide for Data Scientists.** 1. ed. Sebastopol: O'Reilly Media, 2016.

MORDOR INTELLIGENCE. **Residential Real Estate Market in Brazil.** 2023. Disponível em: https://www.mordorintelligence.com/pt/industry-reports/residential-real-estate-market-in-brazil. Acesso em: 07 de abril de 2024.

MIRANDA, Matheus; ZUVIOLLO, Pedro Luiz; PUGLIESI, Jaqueline Brigladori.

Aplicação de Aprendizado de Máquina na Previsão de Preços de Aluguel de Imóveis. Revista de Computação Aplicada, v. 13, n. 1, 2023**.** Disponível em: http://periodicos.unifacef.com.br/resiget/article/view/2764/1849. Acesso em: 14 de Abril de 2024.

NARGESIAN, Fatemeh; SAMULOWITZ, Horst; KHURANA, Umashanthi; KHALIL, Elias B.; TURAGA, Deepak S. **Learning feature engineering for classification.** In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 26., 2017, Melbourne. IJCAI, 2017. p. 2529-2535.

QUESADA, Beatriz. **Dois a cada três brasileiros deixam de negociar imóveis pelo preço, diz Datafolha.** Revista Exame, São Paulo, 20 out. 2024. Disponível em: https://exame.com/mercado-imobiliario/dois-a-cada-tres-brasileiros-deixam-de-negociar-imoveis-pelo-preco-diz-datafolha/. Acesso em: 25 out. 2024.

SOUSA, Maiara Medeiros de. **Modelo Random Forest aplicado a precificação de imóveis à venda em Aracaju, SE**. 2023. Monografia (graduação em Estatística) – Departamento de Estatística e Ciências Atuariais, Centro de Ciências Exatas e Tecnologia, Universidade Federal de Sergipe, São Cristóvão, SE, 2023

TUKEY, John W. **Exploratory Data Analysis**. Reading, MA: Addison-Wesley, 1977.

VAN DER MAATEN, Laurens et al. **Dimensionality reduction: A comparative review**. Journal of machine learning research, v. 10, n. 66-71, p. 13, 2009.

VERLEYSEN, Michel; FRANÇOIS, Didier. **The curse of dimensionality in data mining and time series prediction.** In: CABESTANY, Javier; PRIETO, Antonio; SANDOVAL, Francisco (eds.). Computational Intelligence and Bioinspired Systems. IWANN 2005. Lecture Notes in Computer Science, v. 3512. Berlin: Springer, 2005. p. 437-444. DOI: 10.1007/11494669_93.

WASILEWSKA, Anita. **Preprocessing lecture notes: chapter 3.** 2021. Disponível em: https://www3.cs.stonybrook.edu/~cse634/lecture_notes/07preprocessing.pdf. Acesso em: 24 out. 2024.

WIRTH, Rüdiger; HIPP, Jochen. **CRISP-DM: Towards a standard process model for data mining.** In: INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATIONS OF KNOWLEDGE DISCOVERY AND DATA MINING, 4., 2000, Manchester. Proceedings [...]. Manchester: Practical Application Company, 2000. v. 1, p. 29-39.