UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO CURSO DE ENGENHARIA DE PRODUÇÃO MECÂNICA
CURSO DE ENGENHARIA DE PRODUÇÃO MECÂNICA

Yuri Balczareki Potrich

Precificação de Imóveis em Florianópolis Utilizando Técnicas de Aprendizado de Máquina

Florianópolis
2024

Yuri Balczareki Potrich

Precificação de Imóveis em Florianópolis Utilizando Técnicas de Aprendizado de Máquina

Trabalho de Conclusão de Curso submetida ao curso de Engenharia de Produção Mecânica da Universidade Federal de Santa Catarina para a obtenção do título de Engenheiro Mecânico com habilitação em Engenharia de Produção.
Orientador: Prof. Mauricio Uriona Maldonado, Dr

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Balczareki, Yuri
Precificação de Imóveis em Florianópolis Utilizando Técnicas de Aprendizado de Máquina / Yuri Balczareki ; orientador, Mauricio Uriona Maldonado, 2024.
86 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2024.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. Mercado Imobiliário. 3. Aprendizado de Máquina. 4. Transformação Digital. 5. Estimação de Preços. I. Maldonado, Mauricio Uriona. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

Yuri Balczareki Potrich

Precificação de Imóveis em Florianópolis Utilizando Técnicas de Aprendizado de Máquina

O presente trabalho em nível de trabalho de conclusão de curso foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Mauricio Uriona Maldonado, Dr.Orientador
Universidade Federal de Santa Catarina

Prof. Guilherme Ernani Vieira, Avaliador
Universidade Federal de Santa Catarina

Prof. Ricardo Villarroel Dávalos, Avaliador
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Engenheiro Mecânico com habilitação em Engenharia de Produção.

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 09/07/2024 10:48:53-0300
CPF: ***.104.993-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Monica Mendes Luna,
Dra.Coordenador do Programa
Coordenador do Programa

Documento assinado digitalmente
Mauricio Uriona Maldonado
Data: 09/07/2024 08:54:53-0300
CPF: ***.972.939-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Mauricio Uriona Maldonado, Dr
Orientador

Florianópolis, 03 de Julho de 2024.

AGRADECIMENTOS

Primeiramente, agradeço à minha família e a minha namorada, por me proporcionarem o suporte e apoio necessários para que eu pudesse iniciar e continuar minha jornada acadêmica.

Sou igualmente grato a todas as amizades que fiz na faculdade. As experiências compartilhadas, as noites viradas para terminar um trabalho, o desespero compartilhado antes das provas e as trocas de conhecimento e o apoio mútuo foram essenciais para tornar esta etapa ainda mais significativa e enriquecedora. Agradeço a todos por fazerem parte desta importante fase da minha vida.

Agradeço também ao meu amigo, Victor Rocha Grecco, pelos valiosos conselhos e auxílios em relação à vida acadêmica e profissional. Sua orientação e apoio contínuo foram fundamentais para o meu desenvolvimento pessoal e profssional durante este ciclo.

Por fim, agradeço a todos os professores da universidade, cujos ensinamentos e orientações foram fundamentais para a minha formação como engenheiro. Cada aula, cada conselho e cada incentivo deixaram uma marca indelével na minha trajetória.

RESUMO

A crescente demanda do mercado por modelos de precificação de imóveis eficientes destaca a importância das técnicas de machine learning para a predição de preços, proporcionando aos corretores ferramentas para avaliações mais precisas. Este trabalho objetiva aplicar metodologias de aprendizagem computacional para a avaliação de imóveis, com foco na cidade de Florianópolis, e identificar a metodologia de melhor desempenho. Utilizando dados coletados via web scraping da plataforma Viva Real, o estudo envolveu a análise exploratória e preparação dos dados para treinamento e otimização dos modelos Lasso Regression, Random Forest e XGBoost, sendo este último o mais eficaz, com um R^2 de 0.70, RMSE de 3.67e+05 e MAPE de 24.12%. O estudo contribuiu para avaliar a performance de um modelo genérico para diferentes bairros, além da importância que o mesmo atribui a diferentes amenidades de um imóvel. Conclui-se que a utilização de machine learning é uma abordagem promissora para a precificação de imóveis, com potencial para beneficiar todo o ecossistema imobiliário, mas que necessita de mais pesquisas para melhorar a confiabilidade e a validade da tomada de decisão do algoritmo para então ser utilizada no setor imobiliário. As limitações do estudo e sugestões para trabalhos futuros foram discutidas, apontando para a necessidade de contínua melhoria e adaptação dos modelos com a utilização de um maior volume de dados e de atributos que reflitam melhor as necessidades do consumidor.

Palavras-chave: Mercado Imobiliário; Aprendizado de máquina; Transformação Digital; Estimação de preços

ABSTRACT

The growing market demand for efficient property pricing models highlights the importance of machine learning techniques for price prediction, providing brokers with tools for more accurate evaluations. This work aims to apply computational learning methodologies for property valuation, focusing on the city of Florianópolis, and identify the best-performing methodology. Using data collected via web scraping from the Viva Real platform, the study involved exploratory analysis and data preparation for training and optimizing the Lasso Regression, Random Forest, and XGBoost models, with the latter being the most effective, achieving an R^2 of 0.70, RMSE of 3.67e+05, and MAPE of 24.12%. The study contributed to evaluating the performance of a generic model for different neighborhoods and the importance it assigns to various property amenities. It is concluded that the use of machine learning is a promising approach for property pricing, with the potential to benefit the entire real estate ecosystem. However, more research is needed to improve the reliability and validity of algorithmic decision-making before it can be used in the real estate sector. The study's limitations and suggestions for future work were discussed, pointing to the need for continuous improvement and adaptation of the models by using a larger volume of data and attributes that better reflect consumer needs.

Keywords: Real Estate; Machine Learning; Digital Transformation; Price Estimation

LISTA DE FIGURAS

Figura 1 – Financiamento Imobiliário (SBPE) 12
Figura 2 – Exemplos de modelos com *underfitting, overfitting* e bem ajustados. 20
Figura 3 – Regularização 20
Figura 4 – Bagging 22
Figura 5 – Boosting 23
Figura 6 – Exemplo de estrutura de uma árvore de regressão 24
Figura 7 – Arquitetura do *random forest* 25
Figura 8 – Fluxograma com visão geral do projeto 32
Figura 9 – Processos do Web Scraping. 34
Figura 10 – Portal imobiliário - Viva real 39
Figura 11 – Quantidade de valores nulos por atributo 43
Figura 12 – Distribuição dos dados por macro região 44
Figura 13 – Distribuição dos dados por tipo de anúncio 45
Figura 14 – Distribuição dos dados por bairro 46
Figura 15 – Histograma das variáveis discretas após filtragem. 48
Figura 16 – Matriz de correlação de Spearmn 49
Figura 17 – Histograma para Preço e Área útil 50
Figura 18 – Número de observações para cada Amenidade 51
Figura 19 – Histograma - Preço por m² 53
Figura 20 – Preço médio do m² entre casa e apartamento para cada bairro. . . . 54
Figura 21 – Comparativo do R$/m² entre o FipeZap e Viva Real. 55
Figura 22 – Etapas da limpeza de dados 56
Figura 23 – Separação dos dados em treino e teste 59
Figura 24 – Gráfico de dispersão entre valores reais e valores previstos - Random Forest 63
Figura 25 – Gráfico de dispersão entre valores reais e valores previstos - XGBoost 64
Figura 26 – Gráfico de dispersão entre valores reais e valores previstos - Lasso 65
Figura 27 – Principais variáveis para predição do modelo - XGBoost 69
Figura 28 – R² para diferentes combinações de *features* 70
Figura 29 – MAPE para diferentes combinações de *features* 71
Figura 30 – Modelo regional vs modelo cidade - R². 72
Figura 31 – Modelo regional vs modelo cidade - MAPE 73
Figura 32 – Influência do volume de dados na perfomance do modelo - R² . . . 74
Figura 33 – Influência do volume de dados na perfomance do modelo - MAPE . 75
Figura 34 – Caso de uso 1 76
Figura 35 – Caso de uso 2 76
Figura 36 – Caso de uso 3 76

LISTA DE TABELAS

Tabela 1 – Variáveis extraídas e formatos interpretados pelo Python 40
Tabela 2 – Variáveis escolhidas para análise exploratória 41
Tabela 3 – Análise descritiva: garagem, suites, banheiros e quartos 47
Tabela 4 – Análise descritiva: usableAreas, price, yearlyIptu e monthlyCondoFee 48
Tabela 5 – Remoção de outliers por IQR 49
Tabela 6 – Tabela de Amenidades escolhidas 51
Tabela 7 – Descrição das variáveis 56
Tabela 8 – Valores testados para hiperparâmetros - Random Forest 60
Tabela 9 – Valores testados para hiperparâmetros - XGBoost 61
Tabela 10 – Valores testados para hiperparâmetros - Lasso Regression 62
Tabela 11 – Resumo das métricas para os modelos 65
Tabela 12 – Resumo das métricas por bairro ordenado por R^2. 66

SUMÁRIO

1 INTRODUÇÃO 11
1.1 PROBLEMA DE PESQUISA 11
1.2 OBJETIVOS 13
1.2.1 Objetivo Geral 13
1.2.2 Objetivos Específicos 13
1.3 JUSTIFICATIVAS 14
1.4 ESTRUTURA DO TRABALHO 15
2 FUNDAMENTAÇÃO TEÓRICA 16
2.1 MERCADO IMOBILIÁRIO E SUAS CARACTERÍSTICAS 16
2.1.1 Precificação de imóveis 17
2.1.2 Características Locacionais 17
2.1.3 Características do Próprio Imóvel 18
2.2 MACHINE LEARNING 18
2.2.1 Bias e Variância 19
2.2.2 Validação Cruzada 20
2.2.3 Otimização de hiperparâmetros 21
2.2.4 Modelos de machine learning 21
2.2.4.1 *XGBoost* 21
2.2.4.2 Random Forest 23
2.2.4.3 Regressão Lasso 26
2.2.5 Valores SHAPLEY 27
2.3 MÉTRICAS DE AVALIAÇÃO 28
2.3.1 RMSE 28
2.3.2 Coeficiente de Correlação e (r) R^2 29
2.3.3 MAPE 30
2.4 ESTUDOS PERTINENTES AO TEMA 30
3 METODOLOGIA 32
3.1 COLETA DE DADOS 32
3.1.1 Web Scraping 33
3.2 ANÁLISE EXPLORATÓRIA 35
3.3 PRÉ-PROCESSAMENTO DE DADOS 35
3.3.1 Remoção de duplicatas 35
3.3.2 Remoção de *outliers* 35
3.3.3 Lidando com dados faltantes 36
3.4 *FEATURE ENGINNERING* 36
3.4.1 Tratamento de variáveis categóricas 37
3.5 SEGMENTAR OS DADOS EM CONJUNTO DE TREINO E TESTE . 37

3.6 TREINAMENTO DOS MODELOS 37
3.7 OTIMIZAÇÃO DE HIPERPARÂMETROS 37
3.8 ANÁLISE DOS RESULTADOS 38
4 RESULTADOS 39
4.1 COLETA DE DADOS 39
4.2 PRÉ-PROCESSAMENTO DOS DADOS 42
4.3 SEPARAÇÃO DOS DADOS EM TREINO E TESTE 58
4.4 OTIMIZAÇÃO DE HIPERPARÂMETROS 59
4.5 ANÁLISE DOS RESULTADOS 62
4.6 CASOS DE USO 76
5 CONCLUSÃO 77
REFERÊNCIAS 79
ANEXO A – REPOSITÓRIO PARA O CÓDIGO 84

1 INTRODUÇÃO

Considerado em muitos países como a maior classe de ativos, o mercado imobiliário desempenha um papel fundamental nos sistemas sociais e econômicos. As flutuações nos preços imobiliários têm impactos diretos no sistema financeiro devido ao papel central dos bancos como credores hipotecários e ao uso frequente do imóvel como garantia (PROBST; WRIGHT; BOULESTEIX, 2019).

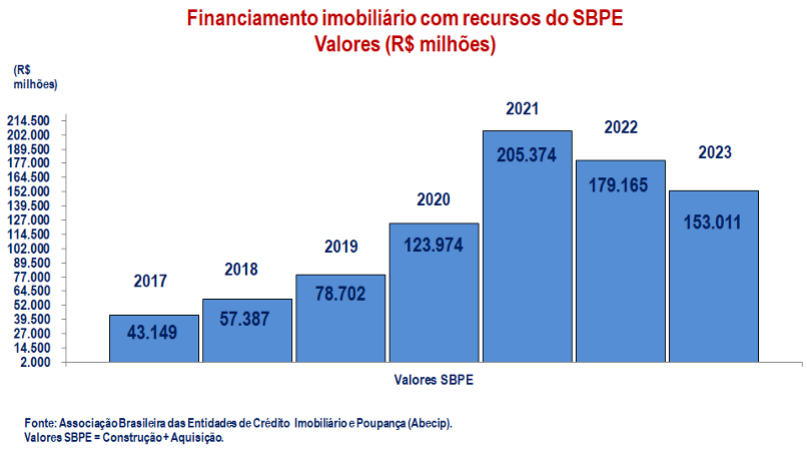
No entanto, adquirir um imóvel é uma operação delicada que requer uma estimativa precisa e objetiva de seu valor antecipadamente. Como a compra de uma casa é a maior transação financeira para a maioria das pessoas (PEDERSEN; WEISSENSTEINER; POULSEN, 2013), conhecer o valor real de uma propriedade é um ativo importante de várias maneiras e permite ao comprador não apenas distinguir entre boas e más negociações, como também negociar efetivamente o preço do imóvel durante a transação. Do lado do vendedor, a estimativa precisa do preço de sua casa antes de colocá-la à venda lhe permite conhecer seu valor de mercado exato. Como resultado, o vendedor pode evitar qualquer risco desnecessário de superestimar ou subestimar o preço de venda. Para Moro (2017) as previsões de preço são usadas no mercado de investimentos para garantir que todo o dinheiro investido seja utilizado de modo que gere o maior retorno.

Deve-se observar que, quando o preço de venda é superestimado, quase certamente causa um atraso na venda, enquanto a subestimação gera uma perda desnecessária de lucro para o vendedor. Além disso, uma estimativa precisa do valor do imóvel é considerada um recurso fundamental para um investidor que deseja diversificar sua carteira devido às alternativas entre títulos imobiliários e outros possíveis investimentos (D'AMATO *et al.*, 2021). Portanto, é crucial e altamente benéfico tanto para vendedores quanto para compradores ter ferramentas que facilitem a estimativa de valores imobiliários.

1.1 PROBLEMA DE PESQUISA

O setor de construção civil desempenha um papel fundamental na economia do Brasil, sendo um dos principais motores de crescimento e desenvolvimento econômico. De acordo com os dados da Associação Brasileira das Entidades de Crédito Imobiliário e Poupança (Abecip) em 2023 os financiamentos imobiliários com recursos da cadertena de poupança (SBPE) totalirazam R$153 bilhões, conforme visto na figura 1.

Figura 1 – Financiamento Imobiliário (SBPE)



Fonte: Indicadores Imobiliários Nacionais - CBIC (2023)

Segundo o boletim econômico da Associação Brasileira de Incorporadoras Imobiliárias (ABRAINC, 2024), o segmento da construção é amplamente reconhecido como um dos principais empregadores da economia. No terceiro trimestre de 2023, o setor gerou 74.217 novos postos de trabalho, representando 13% do total de vagas geradas nacionalmente durante o mesmo período. Conforme os dados da PNAD-C, mais de 7,2 milhões de pessoas estão empregadas nesse setor. O crescimento do emprego também desenha um papel fundamental no impulsionamento do setor imobiliário.

A avaliação de imóveis consiste na determinação do valor de mercado de uma propriedade, definido como o preço mais provável que o mesmo atingiria em uma transação normal, considerando suas características e as condições do mercado naquele momento. No entanto, González (2002) explica que as técnicas empregadas apresentam alguns inconvenientes que resultam na diminuição da precisão destas estimativas, evidenciando a necessidade de aprimoramento.

Apesar de sua significativa contribuição para a economia nacional, o mercado imobiliário enfrenta desafios significativos no que diz respeito à precificação de imóveis. Atualmente o processo de decisão é muito subjetivo e improvisado. Em partes, as dificuldades devem-se principalmente ao desconhecimento dos profissionais sobre o comportamento do mercado, muitas vezes por atuarem simultaneamente em diversas faixas de mercado, em diferentes locais ou com diferentes tipos de imóveis. Geralmente as avaliações são realizadas de forma pontual, usando heurísticas e sem o apoio anterior nem a consolidação posterior em um modelo geral usado no mercado.

Há diversos métodos para avaliação do valor de mercado. Estas são frequentemente realizadas por diversos agentes do mercado, como corretores imobiliários, avaliadores, investidores, gestores de fundos, pesquisadores de mercado e entre outros (MATOS; BARTKIW, 2013). As abordagens mais comuns para modelos de avaliação automatizados baseiam-se em técnicas de regressão paramétrica e não paramétrica.

As técnicas de regressão paramétricas usadas para modelos de avaliação são principalmente baseadas em regressões hedônicas, como análises de regressão linear múltipla (NARULA; WELLINGTON; LEWIS, 2012). Já para os métodos não paramétricos mais utilizados, temos os modelos de aprendizado de máquina, os quais fornecem estimativas rápidas, confiáveis e de baixo custo ao custo de serem uma "caixa-preta", ou seja, difíceis de serem interpretados (VALIER, 2020)

Afim de auxiliar nesse processo decisório dos agentes do ramo, algumas empresas do mercado imobiliário começaram a adotar o uso de inteligência artificial (IA). Segundo Rampini e Re Cecconi (2022) a rotina feita pelos avaliadores, no qual é realizada uma coleta de informações sobre o imóvel, do terreno e da região que o cerca, associando a um valor de venda, podem ser aplicados a um modelo de IA para predizer o valor de venda. Entretanto, caso o conjunto de dados tenha muitas características, o aprendizado do algoritmo pode ser prejudicado, levando o modelo a detectar padrões que não são condizentes com a realidade.

Este trabalho tem como objetivo enriquecer as pesquisas sobre precificação de imóveis. Ao aplicar processos de investigação utilizando modelos de aprendizado de máquina e uma grande quantidade de dados disponíveis publicamente, que refletem a realidade do mercado imobiliário, busca-se propor um modelo de precificação para a região de Florianópolis. Esses dados foram coletados diretamente de um site de uma imobiliária online. Com essa abordagem, pretende-se também caracterizar os principais atributos influentes na precificação dos imóveis e verificar como o modelo de precificação se comporta para diferentes bairros da região, visando disponibilizar os resultados de forma acessível para auxiliar na tomada de decisão de agentes do setor.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Propor um modelo de precificação de imóveis na cidade de Florianópolis/SC utilizando dados de imóveis disponíveis online.

1.2.2 Objetivos Específicos

- Extrair dados do setor imobiliário de uma plataforma online;
- Realizar a análise e tratamento dos dados extraídos;
- Prever o preço de imóveis de venda utilizando modelos de aprendizado de máquina;
- Mapear as características mais relevantes para definição dos valores de compra de imóveis em Florianópolis;

- Mensurar a perfomance do modelo por bairros;

1.3 JUSTIFICATIVAS

O presente trabalho se baseia na crescente demanda do mercado por modelos de precificação de imóveis eficientes e precisos, ressaltando a importância da aplicação de técnicas de machine learning para a predição de preços. Atualmente, tanto corretores quanto investidores imobiliários enfrentam o desafio de dedicar um tempo considerável em pesquisas manuais para estimar o valor de um imóvel, utilizando dados disponíveis em portais imobiliários. Nesse contexto, a proposição de um modelo que possa agilizar e aprimorar esse processo representa uma justificativa fundamental para a realização deste estudo.

Diversos estudos tem sido realizados no que diz respeito a utilização de algoritmos de machine learning para precificação de imóveis. Choy e Ho (2023) em seu estudo procuram demonstrar como modelos de aprendizado podem ter acurácias superiores em relação aos métodos estatísticos tradicionais. Eles concluem que, com informações de preços mais precisas, os compradores podem identificar propriedades que estão com preços muito altos e não valem o investimento, contribuindo para reduzir o desperdício proveniente do desenvolvimento imobiliário desnecessário.

O uso de aprendizado de máquina no mercado imobiliário pode ajudar a aprimorar a tomada de decisões, mitigar riscos e aumentar a eficiência na avaliação, gestão e investimento em propriedades. Em primeiro lugar, algoritmos de aprendizado de máquina podem analisar dados históricos de vendas e outros fatores relevantes, como demografia, localização, tamanho e comodidades, para prever com precisão o valor de uma propriedade (BALDOMINOS *et al.*, 2018).

Ho, Tang e Wong (2021) destacam que os algoritmos avançados de machine learning, como SVM, RF e GBM, são ferramentas promissoras para a predição de preços de imóveis. Apesar de conseguirem previsões com erros significativamente baixos, esses algoritmos possuem limitações, especialmente em termos de seleção cuidadosa de características e interpretação dos coeficientes estimados. Os autores concluem que a aplicação de machine learning na precificação de imóveis ainda está em estágio inicial, e futuras pesquisas podem incorporar dados adicionais de transações imobiliárias de uma área geográfica maior e incluir mais características ou analisar outros tipos de propriedades além de habitações.

É importante salientar que, especialmente para corretores mais novos ou menos experientes, há uma significativa falta de conhecimento e expertise para estimar com precisão os preços dos imóveis. Essa lacuna de conhecimento pode resultar em avaliações subjetivas ou imprecisas, o que, por sua vez, pode impactar negativamente as transações imobiliárias e a satisfação do cliente. Assim, a implementação de modelos de machine learning pode oferecer uma solução promissora para preencher

essa lacuna, fornecendo aos corretores ferramentas e insights objetivos para realizar avaliações mais precisas e fundamentadas.

Dessa forma, além de atender à demanda por eficiência e precisão no mercado imobiliário, a aplicação de técnicas de machine learning também pode desempenhar um papel crucial na capacitação de corretores menos experientes, permitindo-lhes oferecer um serviço mais qualificado e confiável aos clientes. Portanto, a pesquisa neste campo não apenas visa melhorar os processos de precificação, mas também contribuir para o desenvolvimento profissional e aprimoramento das habilidades dos corretores, beneficiando assim todo o ecossistema imobiliário.

1.4 ESTRUTURA DO TRABALHO

O objetivo do primeiro capítulo é fornecer uma breve contextualização do objeto de estudo, destacando os problemas existentes e a motivação inicial. Além disso, são apresentados os objetivos e as justificativas para a realização do trabalho.

No segundo capítulo da fundamentação teórica, são explorados os principais conceitos necessários para o desenvolvimento e estudo do tema abordado. Elabora-se uma breve apresentação da teoria relacionada ao mercado Imobiliário, machine learning e análise de dados.

Em seguida, no capítulo três, são discutidas as metodologias utilizadas no desenvolvimento do projeto de forma sequencial, detalhando os materiais disponíveis para o estudo, os métodos de aprendizado de máquina utilizados, além das técnicas de avaliação e otimização aplicadas. O capítulo finaliza com a apresentação do passo a passo da realização da pesquisa com cada etapa metodológica explicada.

No último capítulo, abordam-se as conclusões abrangentes do estudo, acompanhadas de uma análise dos objetivos inicialmente delineados. Para melhor elucidar as inferências estatísticas, são fornecidas tabelas e gráficos que enfatizam as variáveis cruciais para o modelo. Além disso, são discutidas as considerações finais, avaliando o alcance dos objetivos propostos, identificando limitações e sugerindo direções para futuras investigações.

2 FUNDAMENTAÇÃO TEÓRICA

O capítulo de fundamentação teórica tem como objetivo fornecer uma base sólida de conhecimento e conceitos relevantes para embasar o estudo do presente trabalho. Neste capítulo, são apresentadas as teorias e modelos que sustentam a pesquisa, permitindo uma compreensão aprofundada da avaliação imobiliária e os métodos disponíveis para fundamentação metodológica.

2.1 MERCADO IMOBILIÁRIO E SUAS CARACTERÍSTICAS

O mercado imobiliário desempenha um papel fundamental na economia nacional, devido ao volume expressivo de recursos envolvidos nas transações e à sua relevância social. No entanto, esse mercado apresenta um comportamento distinto em comparação com outros bens de importância econômica. As características singulares dos imóveis tornam a análise de seus valores uma tarefa complexa. A falta de informação por parte dos agentes e o conhecimento limitado sobre os mecanismos de funcionamento do mercado contribuem, para dificultar a análise do mercado imobiliário. Essa combinação de elementos desempenha um papel significativo na explicação das variações de preços observadas nesse mercado (GONZÁLEZ; FORMOSO, 2000).

Dentro do mercado imobiliário, a parte voltada para habitação é aquela que recebe significativamente mais atenção, devido à sua relevância social e ao grande volume de recursos envolvidos. O estudo a seguir se concentra exclusivamente no mercado habitacional.

Os imóveis estão sujeitos às influências dos governos e das economias locais, regionais, nacionais e globais. O poder público exerce uma influência significativa nas mudanças de uso e ocupação do solo, seja por meio de intervenções diretas, como abertura ou alargamento de vias urbanas, ou através do controle ou incentivo à atuação da iniciativa privada, por meio de planos diretores de desenvolvimento. Essas ações podem alterar o comportamento do mercado imobiliário local (BALARINE, 1996).

Para Evans (1995), o mercado imobiliário frequentemente opera de forma informal, tornando desafiador discernir como os agentes adquirem informações sobre as transações. Muitas transações não são devidamente controladas ou registradas, devido a questões como evasão fiscal ou encargos de transferência, que podem representar custos substanciais. Assim, a falta de transparência caracteriza esse mercado.

Além disso, Abramo (1988) explica que o mercado também é dinâmico, embora tais variações sejam relativamente lentas. A realização de obras, tais como escolas, parques, avenidas, hospitais, universidades, *shopping centers* ou indústrias, introduz modificações ao seu redor e também a uma área de maior abrangência de sua vizinhança.

Tais bens são heterogêneos por natureza, pois cada imóvel terá quantias dife-

rentes de cada um dos atributos valorizados pelo consumidor. Por isso, são comumente chamados de “bens compostos”, visto que a comparação entre eles exige uma ponderação dos diversos atributos de interesse (ROBINSON, 1979).

2.1.1 Precificação de imóveis

Existem três métodos tradicionais de avaliação imobiliária: o método de comparação de vendas, o método de renda e o método de custo (LINNE; KANE; DELL, 2000).

Segundo o método de comparação de vendas, que consiste em avaliar habitações disponíveis no mercado com base nas suas características intrínsecas (número de quartos, banheiros, garagem, suítes, área total, etc.) e extrínsecas (distância até pontos de interesse, como hospitais, escolas, parques e metrô), o valor é ajustado de acordo com as diferenças, uma vez que os imóveis apresentam variações (ARRAES; SOUSA FILHO, 2008).

De acordo com Fiker (2001), o método da renda consiste em descobrir o valor do imóvel através da capitalização de sua renda líquida real ou prevista. A abordagem é utilizada para avaliar imóveis que geram renda, como estabelecimentos comerciais, prédios com escritórios ou apartamentos alugados, utilizados para serviços ou produção.

Por fim, no método de abordagem de custo, o valor do imóvel é determinado pelos custos de construção menos a depreciação. Esta abordagem costuma ser aplicada apenas a edifícios, e é muito adequada para escolas, objetos de infraestrutura de engenharia e similares, que não geram renda e para os quais há apenas alguns para comparar.

2.1.2 Características Locacionais

Para González (2002), uma das características cruciais do mercado imobiliário se dá pela sua imobilidade de oferta e, ao mesmo tempo, pela demanda ser localizada espacialmente. A demanda por imóveis é determinada por condições específicas de cada localidade, levando em consideração padrões de renda, níveis de emprego, preferências da população e outros fatores. Essa demanda varia significativamente de um local para outro, refletindo as particularidades de cada região. Além disso, cada localidade provê diferentes ofertas de serviços públicos, externalidades negativas e amenidades, entregando diferentes níveis de qualidade (ou desejabilidade).

O valor de localização de um imóvel está associado à sua acessibilidade, incluindo a oferta e qualidade das vias e meios de transporte, bem como às características da vizinhança e ao uso do solo nas proximidades. No entanto, a mensuração desses efeitos é desafiadora, uma vez que não podem ser diretamente quantificados. Geralmente, são utilizadas variáveis substitutas, como a renda média da população ou

a distância em relação ao centro comercial-histórico da área urbana, com o intuito de estimar esses impactos (GONZÁLEZ; FORMOSO, 2000).

Os efeitos que a vizinhança exerce sobre o preço são igualmente importantes, porém de difícil mensuração. Din, Hoesli e Bender (2001) trazem em seus estudos os diversos fatores que contribuem para esse efeito, tais como padrão dos imóveis na vizinhança (no ambiente construído), o grau de escolaridade e a renda média dos residentes, a qualidade do ar, a disponibilidade de escolas e transporte público (ônibus e metrôs).

Para Can (1998), vizinhanças podem ser definidas como entidades espaciais discretas ("áreas físicas") que contêm domicílios e estruturas de habitação com características semelhantes. Tipicamente, os domicílios apresentam características sociais, econômicas e demográficas similares dentro dos bairros. Segundo o autor, espera-se que os preços dos imóveis alterem-se sistematicamente ao longo da área urbana. Estas variações são também ditas contínuas, ou seja, os valores não aparecem de forma aleatória. Logo, podem ser mapeados a partir de dados do mercado, seja usando abordagens que levem em conta o geoprocessamento ou através de superfícies matemáticas

2.1.3 Características do Próprio Imóvel

Existe uma enorme variedade de produtos no mercado imobiliário, grande parte deles se deve aos aspectos ligados ao próprio imóvel e à tecnologia por trás de sua construção. Os imóveis apresentam diferenças em idade, qualidade de construção, tamanho e diversos outros elementos, os quais se refletem nas variações de preços no mercado. Tamanha heterogeneidade dos imóveis e de suas localizações, segundo González e Formoso (2000), dificulta a comparação de preços, pois a informação sobre os vários atributos que compõem o produto nem sempre está disponível aos agentes.

2.2 MACHINE LEARNING

A inteligência artificial (IA) é um subcampo da ciência da computação e concentra-se no *design* de programas de computador e máquinas capazes de realizar tarefas nas quais os seres humanos são naturalmente bons, incluindo compreensão de linguagem natural, compreensão de fala e reconhecimento de imagens. Por volta do século XX, a aprendizagem de máquina surgiu como um ramo da IA, proporcionando uma nova abordagem para o *design* da IA, baseada em uma compreensão conceitual de como o cérebro humano funciona.

A utilização de IA e ML no mercado imobiliário apresenta um cenário semelhante. A regressão de preço hedônico se estabeleceu como a principal abordagem na estimativa de preços e aluguéis, por conseguir tratar um ativo imobiliário como a

soma de suas características individuais. Embora os modelos de ML tenham se mostrado úteis na modelagem de preços hedônicos para fins preditivos de imóveis, tais modelos carecem de transparência e na revelação das relações teóricas subjacentes (MULLAINATHAN; SPIESS, 2017).

Escolher um algoritmo de regressão apropriado para uma tarefa específica requer prática e experiência, pois cada algoritmo possui suas próprias peculiaridades e é baseado em certas suposições. Segundo o teorema *"no free lunch"* proposto por Wolpert (1996), nenhum modelo único funciona melhor em todos os cenários possíveis. Na prática, recomenda-se comparar o desempenho de pelo menos alguns algoritmos de aprendizado diferentes, para selecionar o melhor modelo para o problema específico; esses podem diferir no número de características ou exemplos, na quantidade de ruído em um conjunto de dados e se as classes são linearmente separáveis.

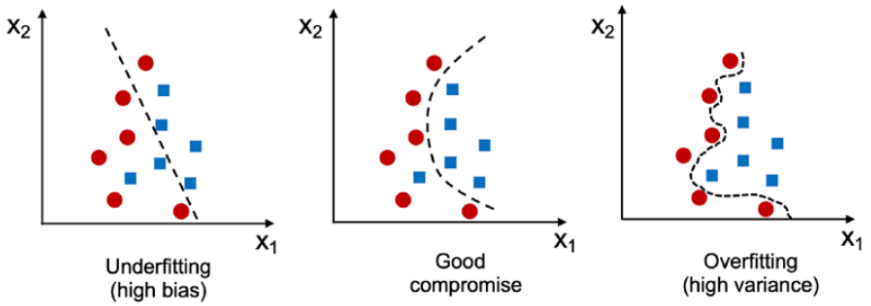
Para Raschka, Patterson e Nolet (2020), os cinco principais passos envolvidos no treinamento de um algoritmo de aprendizado de máquina supervisionado podem ser resumidos na seguinte forma:

1. Selecionar características e coletar exemplos de treinamento rotulados.
2. Escolher uma métrica de desempenho.
3. Escolher um algoritmo de aprendizado e treinar um modelo.
4. Avaliar o desempenho do modelo.
5. Alterar as configurações do algoritmo e ajustar o modelo.

2.2.1 Bias e Variância

O controle da flexibilidade de um algoritmo de *machine learning* é fortemente ligado ao equilíbrio entre viés e variância. O viés está relacionado à capacidade do modelo de se ajustar adequadamente aos dados, ou seja, se ele consegue capturar com precisão a relação entre as variáveis independentes e dependentes (Fig 2). Por outro lado, a variância está relacionada à sensibilidade do modelo às flutuações nos dados de treinamento, ou seja, o quão suscetível ele é a pequenas variações nos dados de treino (GOLDSTEIN; NAVAR; CARTER, 2017).

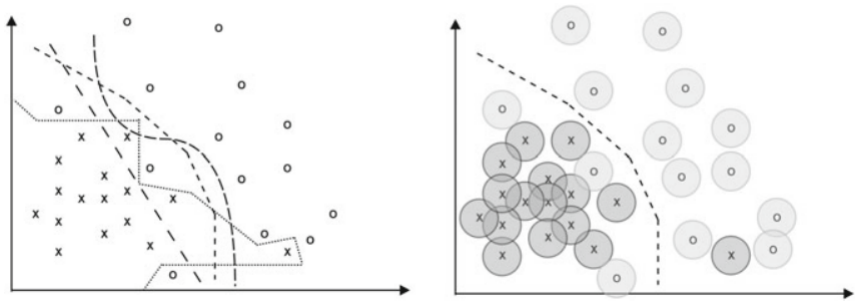
Figura 2 – Exemplos de modelos com *underfitting, overfitting* e bem ajustados.



Fonte: (RASCHKA; PATTERSON; NOLET, 2020)

O processo de redução da variância de um modelo é conhecido como Regularização 3. Esse processo visa ajustar o modelo de forma a capacitar sua capacidade de generalização para novos dados que não foram utilizados no treinamento.

Figura 3 – Regularização



Fonte: (SKANSI, 2018)

2.2.2 Validação Cruzada

A validação cruzada (*cross validation*) é uma das técnicas mais amplamente utilizadas para avaliar diferentes métodos ou parâmetros em algoritmos de *machine learning*. Ela envolve dividir aleatoriamente o conjunto de dados de treinamento em k partes de tamanho igual, onde $k-1$ partes são utilizadas para treinar o modelo e a parte restante é reservada para a avaliação da sua performance. Esse processo é repetido várias vezes, até que todas as partes tenham sido utilizadas tanto para treinamento quanto para avaliação, resultando em k estimativas de perfomance (KUHN; JOHNSON *et al.*, 2013). Repetições desse procedimento podem ser realizadas para melhorar a precisão das estimativas, permitindo uma avaliação robusta e geral do desempenho do modelo.

Ainda para Kuhn, Johnson *et al.* (2013), a escolha do parâmetro k na validação cruzada não segue uma regra precisa, embora sejam comuns divisões em 5 ou 10 partes. À medida que aumentamos o valor de k, a diferença de tamanho entre o conjunto de treinamento original e os subconjuntos reamostrados diminui, reduzindo assim o viés da técnica de validação cruzada. Entretanto, um aumento em k também implica em um aumento no tempo necessário para obter o resultado final da validação cruzada.

2.2.3 Otimização de hiperparâmetros

Para desenvolver um modelo de *machine learning* eficaz, é fundamental explorar várias opções. A seleção dos hiperparâmetros desempenha um papel crucial na definição da arquitetura ideal do modelo, pois esses parâmetros são responsáveis por configurar diversos aspectos de um algoritmo de aprendizado. Eles têm um amplo impacto no resultado final do modelo, especialmente em casos de modelos baseados em árvores de decisão, que possuem uma grande variedade de parâmetros. A otimização desses hiperparâmetros ajuda a encontrar um equilíbrio adequado entre viés e variância, maximizando assim o desempenho do modelo (SHARMA; HARSORA; OGUNLEYE, 2024).

Na prática, é necessário ajustar continuamente os hiperparâmetros e treinar diversos modelos com diferentes combinações de valores e, em seguida, comparar o desempenho dos modelos para escolher o melhor. Na busca por hiperparâmetros otimizados, várias abordagens foram desenvolvidas pela literatura, sendo uma das mais empregadas o *grid search*. Esta técnica segue uma estratégia exaustiva, testando todas as combinações possíveis de valores de hiperparâmetros dentro de um intervalo pré-definido pelo usuário em um conjunto de validação cruzada (WU *et al.*, 2019). Utilizar esse método permite explorar uma ampla variedade de configurações e encontrar aquela que resulta no melhor desempenho do modelo de acordo com as métricas avaliativas. No entanto, é importante ressaltar que a busca em grade requer grande capacidade computacional, pois testa múltiplas combinações diferentes para o mesmo modelo, o que pode gerar sobrecarga e lentidão na busca dos hiperparâmetros.

2.2.4 Modelos de machine learning

2.2.4.1 *XGBoost*

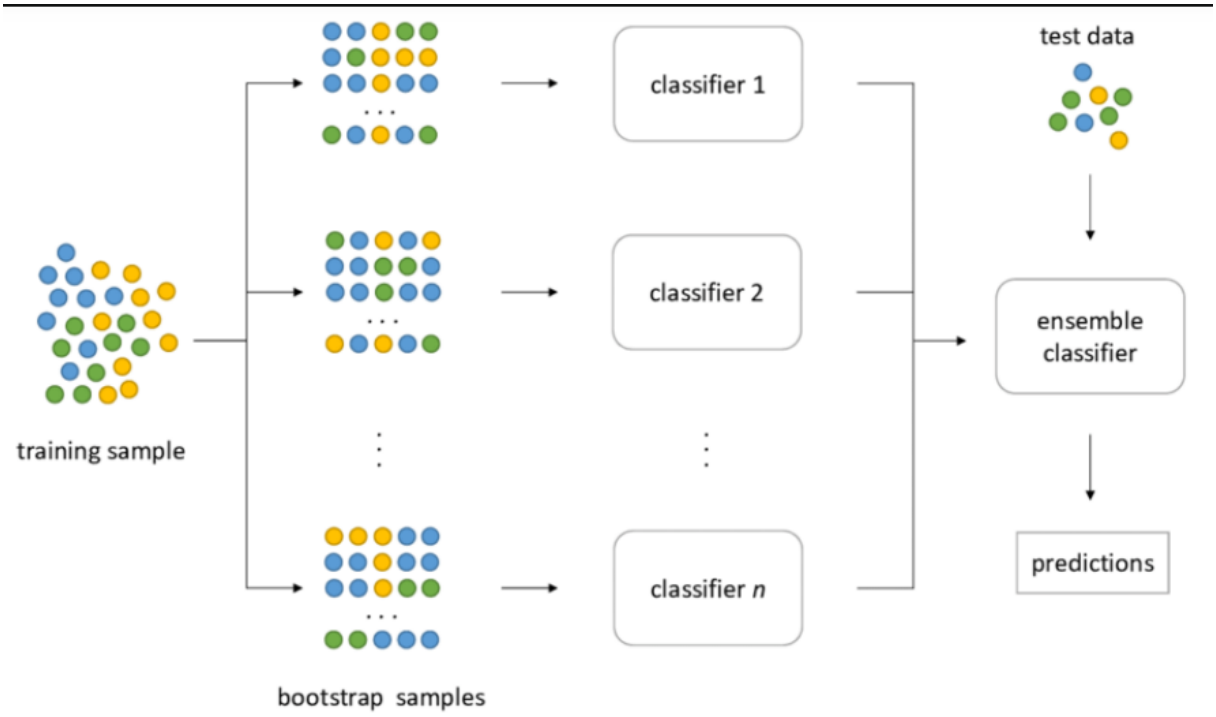
De acordo com Chen e Guestrin (2016), o *XGBoost* é um modelo de integração de impulsionamento que combina o poder do algoritmo de *gradient boosting* com a flexibilidade das árvores de decisão. Ele utiliza várias árvores de decisão, conhecidas como aprendizes fracos, para realizar a tarefa de aprendizado. Além dessas vantagens, o modelo possui ainda uma funcionalidade de validação cruzada incorporada que

permite avaliar o desempenho do modelo ao longo do processo de treinamento e ajustar os hiperparâmetros de acordo com a necessidade.

Ao invés de usar o método tradicional de procura, o XGBoost utiliza diretamente os valores das primeiras e segundas derivadas da função de perda, melhorando o desempenho do algoritmo via técnicas de pré-ordenamento e número de bits dos nós. Após a regularização, o modelo XGBoost escolhe o modelo mais simples que atenda ao desempenho esperado. O resultado final é um modelo capaz de realizar previsões precisas, mesmo que os modelos individuais não sejam tão bons. Graças a essa regularização, o algoritmo consegue suprimir o *overfitting* dos aprendizes fracos em cada iteração, garantindo que apenas os fatores relevantes sejam considerados na construção das árvores de decisão (LI *et al.*, 2021).

Bagging (figura 4 envolve a criação de múltiplos conjuntos de dados de treinamento através da amostragem com substituição (*bootstrap*) do conjunto de dados original. Cada conjunto é utilizado para treinar um modelo diferente e independente. Os resultados dos modelos são combinados geralmente pela média (para regressão) ou por votação (para classificação), reduzindo assim a variância geral do modelo *ensemble*.

Figura 4 – Bagging

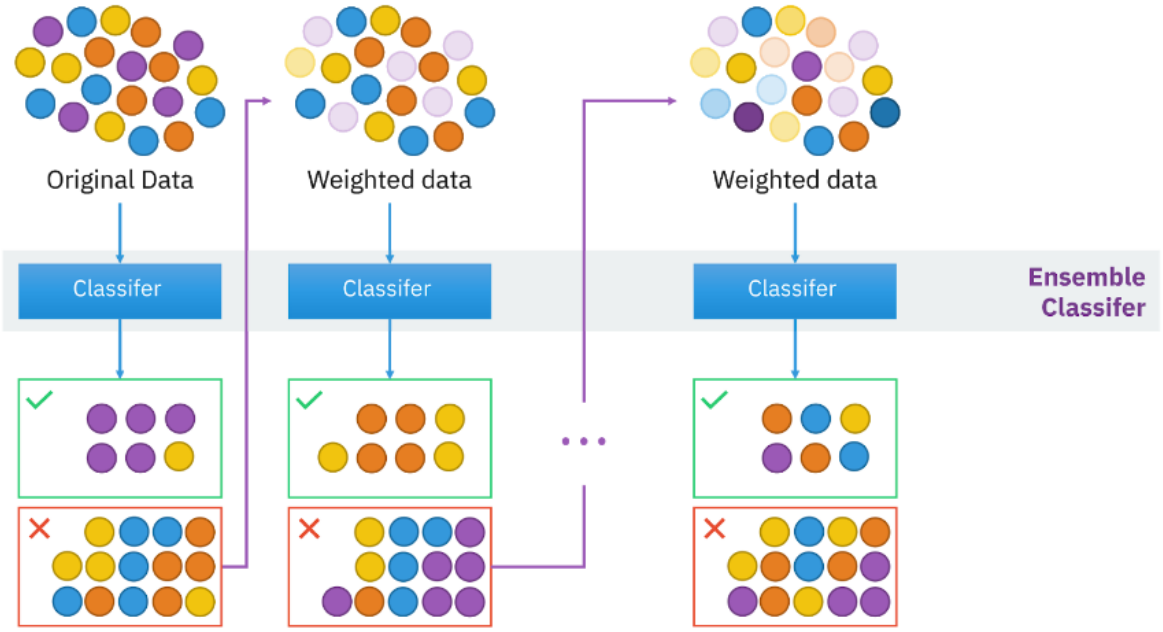


Fonte: (HACHCHAM, 2024)

Já o boosting (5 é uma técnica de *ensemble* que utiliza múltiplos modelos de forma sequencial. Cada modelo é treinado para corrigir os erros dos modelos anteriores na sequência, dando mais peso aos exemplos de treinamento que foram

mal previstos anteriormente. Esse ajuste progressivo permite que cada novo modelo se concentre nas regiões onde os modelos anteriores tiveram dificuldades, melhorando gradualmente a precisão do conjunto final. No XGBoost, por exemplo, são construídas árvores de decisão iterativamente, onde cada nova árvore é ajustada para minimizar os resíduos deixados pelos modelos anteriores, utilizando uma abordagem de gradient boosting para otimizar o processo de aprendizado e predição.

Figura 5 – Boosting



Fonte: (PROSKURIN, 2024))

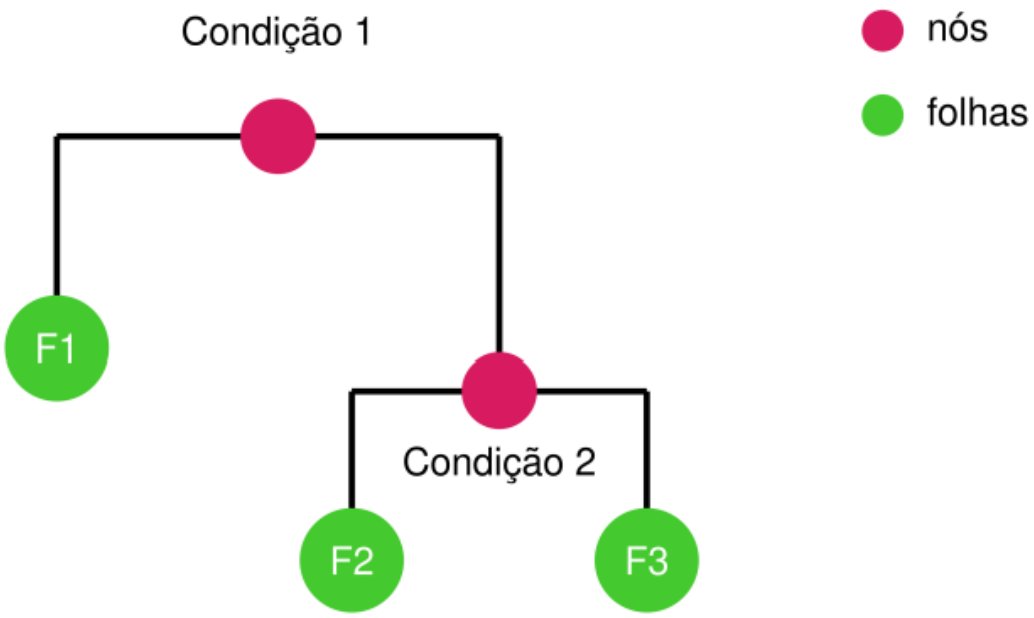
A maior vantagem do XGBoost é que as características dos dados são ordenadas antes do treinamento e armazenadas em blocos. Como resultado, os blocos existentes podem ser utilizados para iterações subsequentes, o que reduz significativamente a quantidade de cálculos necessários. O XGBoost gera valores que medem a importância das características, permitindo identificar e quantificar o peso de cada uma das características em seu conjunto de dados, o que pode reduzir o risco de *overfitting* do modelo (CHEN; GUESTRIN, 2016).

2.2.4.2 Random Forest

O Random Forest é um método de aprendizado de conjunto modificado a partir das árvores de decisão (BREIMAN, 2001). Uma árvore de decisão é um método de classificação não paramétrico que recursivamente particiona o conjunto de dados de treinamento até que cada partição individual consista inteiramente de exemplos de uma classe. A árvore é composta por folhas e nós de decisão. As folhas representam

as classes e ficam no extremo dos ramos. Cada nó contém uma divisão, a qual consiste em um teste aplicado a um ou mais atributos, gerando um ramo (e assim uma sub-árvore) para cada resultado possível do teste (IZBICKI; SANTOS, 2020).

Figura 6 – Exemplo de estrutura de uma árvore de regressão



Fonte: (IZBICKI; SANTOS, 2020))

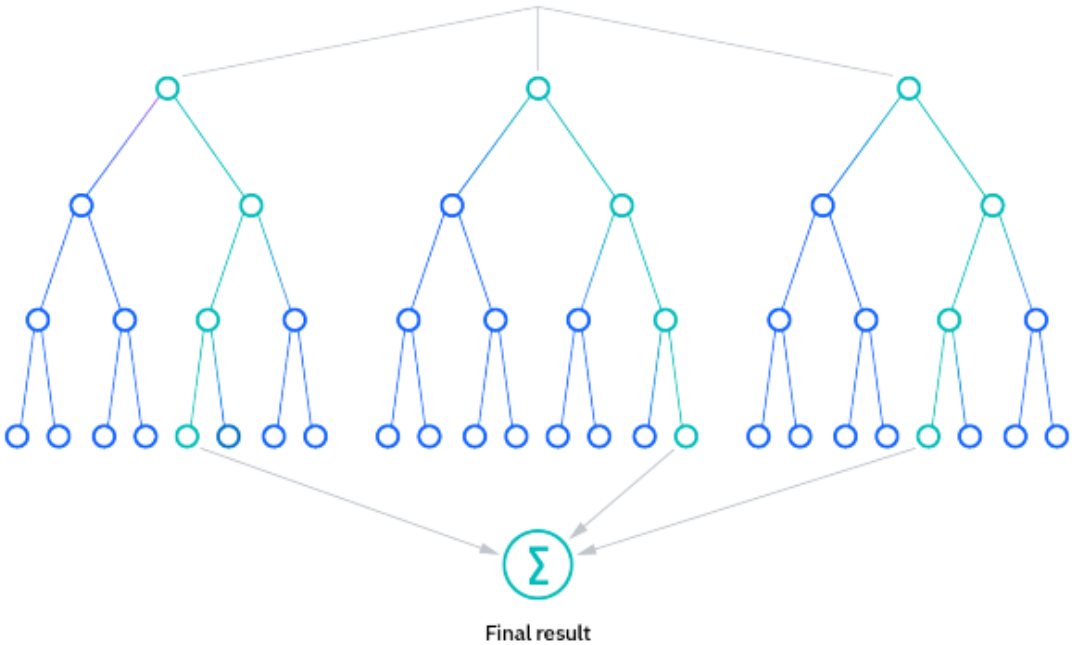
A utilização de uma árvore para prever uma nova observação segue o seguinte procedimento: começamos pelo topo da árvore e verificamos se a condição descrita no nó inicial é satisfeita. Se sim, seguimos para o nó à esquerda; caso contrário, seguimos para o nó à direita. Esse processo é repetido até chegarmos a uma folha da árvore. No exemplo ilustrativo da Figura 7. e a condição 1 for satisfeita, a predição é feita com base na informação fornecida pelo nó correspondente, identificado como F1. Se a condição não for satisfeita, seguimos para a direita e encontramos outra condição. Se essa segunda condição for satisfeita, a observação é prevista como F2; caso contrário, é prevista como F3.

As árvores dividem os dados de treinamento em subconjuntos distintos, cada um descrito por uma regra simples a respeito de um ou dois atributos. Uma das principais vantagens das árvores de decisão é que o modelo é muito poderoso, o mesmo retorna de maneira compreensiva a importância atribuída para cada variável independente na hora de realizar as predições.

Uma característica notável das árvores de regressão é sua alta interpretabilidade. No entanto, em comparação com outros estimadores, elas tendem a ter um poder preditivo relativamente baixo. Para contornar essa limitação, métodos como *bagging* e florestas aleatórias (BREIMAN, 2001) são utilizados. Esses métodos combinam várias árvores para fazer previsões para um mesmo problema, melhorando assim a precisão do modelo.

Os métodos de *bagging* e florestas aleatórias consistem em criar N árvores distintas e combinar seus resultados para melhorar o poder preditivo em relação a uma árvore individual 7. Embora o método de Bagging produza preditores que não são tão facilmente interpretáveis quanto as árvores, ele permite a criação de uma medida de importância para cada variável explicativa. Essa medida de importância é calculada com base na redução da soma dos quadrados dos resíduos (RSS - *residual sum of squares*) de cada divisão. Calcula-se o total de redução da soma dos quadrados dos resíduos para todas as divisões feitas com base nessa variável. Tal processo é repetido para todas as árvores, e então é calculada a média dessa redução. É graças a essa média da redução que se consegue medir a importância de cada variável (IZBICKI; SANTOS, 2020).

Figura 7 – Arquitetura do *random forest*



Fonte: (IBM, 2024))

Para a construção do modelo é necessário a otimização de três parâmetros

- *ntree*: Indica a quantidade de árvores a serem criadas em um modelo de*random*

forest. Não existe um método universal para definir o número ideal de árvores a serem construídas; geralmente, bases de dados menores e mais simples requerem menos árvores, enquanto bases maiores e mais complexas demandam mais. É geralmente entendido que um maior número de árvores tende a melhorar o desempenho do modelo. No entanto, isso também pode aumentar a tendência ao *overfitting* e o custo computacional (IZBICKI; SANTOS, 2020).

- *mtry*: Responsável pela seleção aleatória do número de variáveis que serão consideradas para a criação dos nós. Modelos com um alto valor de "mtry"tendem a gerar árvores mais diversificadas, o que pode levar ao *overfitting*. Por outro lado, aqueles com valores mais baixos tendem a explorar com mais profundidade as variáveis que têm um efeito moderado na variável resposta. Entretanto, os mesmos podem resultar em árvores de baixo desemppenho, uma vez que são construídas com base em um conjunto menor de variáveis, o que pode resultar na seleção de variáveis importantes (PROBST; WRIGHT; BOULESTEIX, 2019).

- *nodesize*: Consiste no tamanho mínimo que um nó de uma árvore deve ter, controlando assim o tamanho geral da árvore. Estimar valores baixos para o "nodesize"resulta em um menor custo computacional e árvores mais profundas, ou seja, com mais divisões até os nós finais. Por outro lado, estimar valores altos leva a árvores mais custas e propensas ao *overfitting* (PROBST; WRIGHT; BOULESTEIX, 2019).

2.2.4.3 Regressão Lasso

A regressão Lasso (*Least Absolute Shrinkage and Selection Operator*) é uma extensão da regressão linear e também um método de regularização, isto é, métodos que adicionam uma penalização na equação dos minímos quadrados. O mesmo introduz uma penalidade nos coeficientes do modelo, com o objetivo de reduzir a variância das estimativas e realizar a seleção de variáveis, como demonstrado por (TIBSHIRANI, 1996). Dessa forma, o lasso é capaz de, simultaneamente, estimar os parâmetros e selecionar covariáveis para o modelo.

Para Hastie, Tibshirani e Wainwright (2015), o método LASSO emerge como uma opção intrigante para a seleção de variáveis, especialmente útil em análises de grandes conjuntos de dados, particularmente quando o número de variáveis é maior do que o número de observações. O LASSO também assegura que muitos coeficientes associados a essas variáveis sejam nulos, explicitando apenas as características mais relevantes para o estudo do problema.

A técnica minimiza a soma dos quadrados dos resíduos do modelo usando um parâmetro de ajuste (*tuning parameter*), que deve ser maior do que a soma dos valores absolutos dos coeficientes do modelo, resultando em uma penalização do tipo L1. Isso

leva vários coeficientes a serem forçados a zerar, originando o termo *"shrinkage"*. Os parâmetros não nulos podem ser considerados os mais significativos na construção do modelo, tornando assim a técnica útil para seleção de variáveis, explicando o termo *"selection"* na sigla.

Dado um conjunto de dados X com n observações e p características, e um vetor de respostas y, o objetivo do algoritmo de regressão Lasso é encontrar o vetor de coeficientes β que minimiza a função de perda com uma penalização L1, representada por:

$$\min_{\beta} \left(\frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right) \tag{1}$$

Onde:

- $\|y - X\beta\|_2^2$ é a soma dos quadrados dos resíduos do modelo de regressão linear.
- $\|\beta\|_1$ é a norma L1 dos coeficientes do modelo, que é a soma dos valores absolutos dos coeficientes.
- λ é o parâmetro de ajuste *(tuning parameter)*, que controla o balanço entre o ajuste do modelo e a complexidade (esparsidade) dos coeficientes. Um valor maior de λ leva a uma maior penalização, resultando em mais coeficientes sendo forçados a zero.

2.2.5 Valores SHAPLEY

Os valores de Shapley, propostos por Shapley *et al.* (1953), surgiram como uma solução para o problema de jogos cooperativos na Teoria dos Jogos. Esses valores visam representar a contribuição específica de cada jogador de uma coalizão para a obtenção do payoff, possibilitando a alocação justa dos ganhos entre os jogadores. Esses valores se fundamentam em quatro axiomas fundamentais:

1. Simetria: se dois agentes contribuem com o mesmo valor para todas as possibilidades de coalizões, eles deveriam receber a mesma proporção do payoff (e ter o mesmo valor de Shapley).
2. Jogador Dummy: se o agente não contribui com nenhum valor para nenhuma coalizão, ele não deverá receber nada (terá um valor de Shapley igual a 0).
3. Aditividade: ao dividir um jogo em duas partes, o payoff do jogo original deve ser igual à soma dos payoffs dos jogos separados.
4. Eficiência: a soma dos valores de Shapley de todos os agentes é igual ao valor total do jogo.

A utilização dos valores de Shapley para ajudar na interpretação de modelos de *machine learning* surge da representação do problema de modelagem como um jogo cooperativo (LUNDBERG; LEE, 2017). Nesse cenário, para uma observação específica, a previsão do modelo desempenharia o papel de *payoff*, enquanto os regressores agiriam como diferentes agentes, cada um contribuindo para o resultado final. Assim, o valor de Shapley de cada variável quantifica o quanto as características individuais contribuem para o desempenho do modelo em um conjunto de dados.

Os valores de Shapley têm sido estudados como uma medida de importância de características em vários contextos. Usando essas estimativas de importância, as características podem ser classificadas e selecionadas ou removidas conforme necessário. Esta abordagem tem sido aplicada a várias tarefas que auxiliam na interpretação e seleção de características para melhorar a performance do modelo (ROZEMBERCZKI *et al.*, 2022).

2.3 MÉTRICAS DE AVALIAÇÃO

2.3.1 RMSE

O *Root Mean Square Error* é uma métrica comumente usada em problemas de regressão. Essa medida calcula o erro quadrático médio entre as observações reais e as previsões do modelo para todo o conjunto de dados, e é definida na Equação (2) como:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Elementos da fórmula do RMSE:

- n : Número de pontos de dados
- y_i : Valor real (observado) do i-ésimo ponto de dado
- $\hat{y}_i$: Valor previsto do i-ésimo ponto de dado

A suposição subjacente ao usarmos o RMSE é que os erros são imparciais e seguem uma distribuição normal. Além disso, esta métrica é facilmente afetada pela presença de outliers nos dados (CHAI; DRAXLER, 2014).

Mesmo que o RMSE seja comumente a medida de desempenho mais usada nas tarefas de regressão, para alguns contextos, pode ser preferível usar outra função. Para os casos que não foi possível a remoção de outliers, pode-se considerar o uso do *Mean Absolute Error* (às vezes também chamado de *Average Absolute Deviation*) (GÉRON, 2017), dado pela equação Equação (3)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

– n : Número de pontos de dados

– y_i : Valor real (observado) do i-ésimo ponto de dado

– $\hat{y}_i$: Valor previsto do i-ésimo ponto de dado

Tanto o RMSE quanto o MAE são formas de medir a distância entre dois vetores: o vetor de previsões e o vetor de valores-alvo.

2.3.2 Coeficiente de Correlação e (r) R²

Garson (2009) afirma que o coeficiente de correlação de Pearson r é uma medida de associação bivariada (força) do grau de relacionamento entre duas variáveis, ou seja, o grau de associação linear entre variáveis. Outra métrica relevante para a avaliação de um modelo é o coeficiente de múltipla determinação R^2, uma ferramenta estatística que aponta a porcentagem de variação em um determinado atributo alvo que pode ser explicada pelo modelo. O valor de R-quadrado serve como indicador de quão eficaz o modelo está ajustado e da precisão de suas previsões em amostras de dados não observadas. Pode ser definido pela Equação (4) como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Onde:

– r : Coeficiente de correlação de Pearson

– n : Número de pontos de dados

– x_i : Valor do i-ésimo ponto de dado na primeira variável

– y_i : Valor do i-ésimo ponto de dado na segunda variável

– $\bar{x}$: Média (valor médio) dos valores na primeira variável

– $\bar{y}$: Média (valor médio) dos valores na segunda variável

Para Craven e Islam (2011), tal coeficiente expressa a quantidade de variação na variável alvo a qual pode ser explicada pela variação nos atributos do conjunto de dados. No entanto, a análise isolada desse coeficiente não é suficidente para determinar validar um modelo, pois é possível que um modelo ruim obtenha um alto valor de R^2. Sua utilidade está em comparar dois modelos válidos.

2.3.3 MAPE

O *Mean Absolute Percentage Error* (MAPE) é uma medida de erro relativo que utiliza valores absolutos para evitar que os erros positivos e negativos se anulem mutuamente e utiliza erros relativos para permitir a comparação da precisão da previsão entre modelos de séries temporais.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \tag{5}$$

- y_i: Valor real observado.
- $\hat{y}_i$: Valor previsto pelo modelo.
- n: Número total de observações.

Um valor de MAPE baixo significa que as previsões estão próximas dos valores observados (O modelo está acertando bem), enquanto um valor alto indica que as previsões estão longe dos valores observados (o modelo precisa ser melhorado).

2.4 ESTUDOS PERTINENTES AO TEMA

Kok, Koponen e Martínez-Barbosa (2017) descobriram vantagens na aplicação de técnicas de *automated valuation models* (AVM), incluindo uma maior precisão e velocidade de análise em relação aos modelos tradicionais. Um AVM é um serviço que utiliza modelos matemáticos para fornecer o valor estimado de um imóvel para um dado ponto específico no tempo. Em seus estudos, utilizaram-se o modelo de árvore de regressão com métodos de *random forest* aliados ao *gradient boost* em um grande conjunto de dados de transações residenciais. Nos modelos, incluíram-se milhares de variáveis de diversas outras fontes para obter previsões com um erro absoluto de 9%, em comparação com um erro estimado de 12% nas metodologias tradicionais de avaliação manual. Os resultados também mostraram uma melhoria significativa nas previsões em relação aos modelos tradicionais de regressão hedônica quando variáveis importantes, como a receita líquida de operação (RLO), estão ausentes. Outro achado interessante é que o uso de conjuntos de dados maiores que incluíam várias regiões geográficas realmente aprimorou os resultados das previsões do RLO, mostrando que a relação entre variáveis explicativas, como distância para o transporte público e presença de escolas, se mantém entre as regiões (KOK; KOPONEN; MARTÍNEZ-BARBOSA, 2017).

No estudo realizado por Guan *et al.* (2014), o sistema de inferência difusa *adaptive neuro-fuzzy inference system* (ANFIS) foi utilizado para avaliação imobiliária. Eles empregaram um banco de dados com 20.192 registros de vendas para uma região

do meio-oeste dos Estados Unidos, coletados entre 2003 e 2007. As características usadas no modelo foram localização, ano de construção e venda, metragem quadrada em cada andar (incluindo porão e garagem), número de banheiros, número de lareiras, presença de ar condicionado central, tipo de lote e tipo de construção. Sua abordagem utilizou redes neurais com inferência difusa, na qual se observou a superioridade do ANFIS em relação à *multiple regression analysis* (MRA). Os mesmos destacaram que o uso apenas da localização em vez de todas as variáveis pode melhorar os resultados (ALMAQBALI *et al.*, 2019).

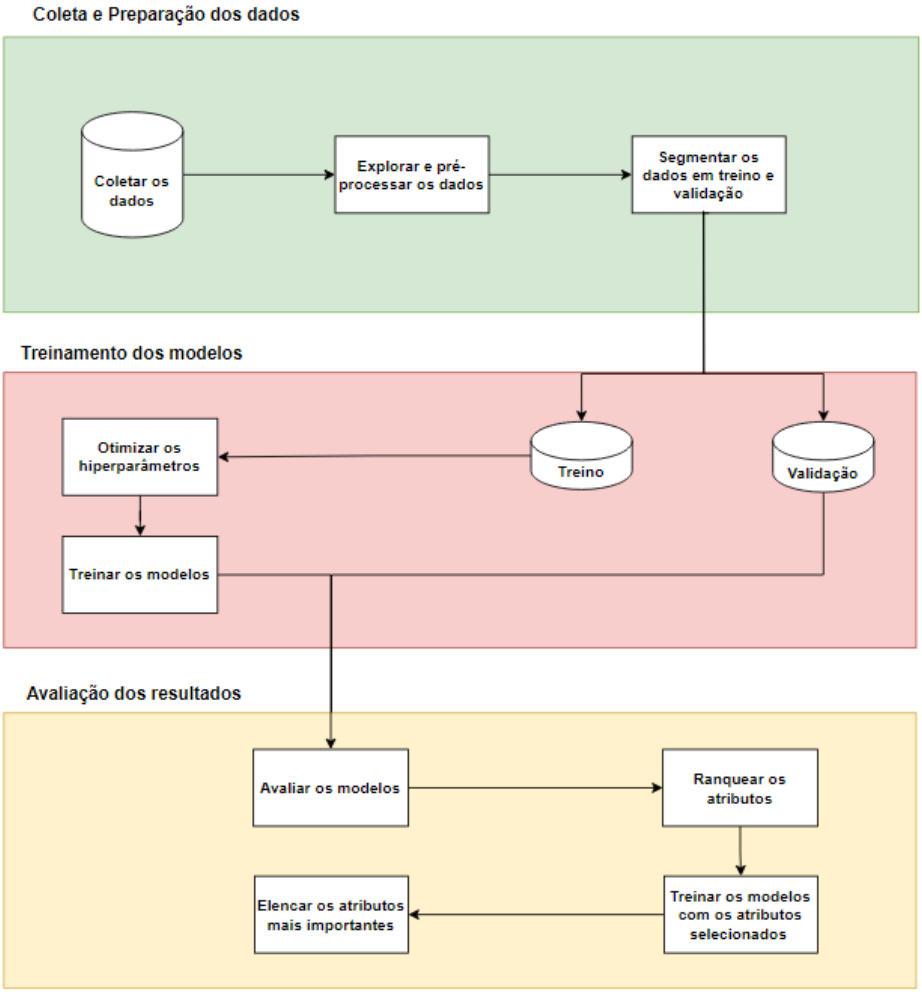
Alguns autores têm utilizado técnicas de aprendizado de máquina para prever o preço de ativos imobiliários individuais. Um exemplo é o estudo de Park e Bae (2015), que analisaram dados de moradias de 5359 casas na região de Virgínia, combinados de várias bases de dados entre 2004 a 2007. Esta base de dados continha 76 atributos, dos quais apenas 28 foram selecionados após aplicação de filtros utilizando o teste t e regressão logística. Dezesseis desses atributos são características físicas, referentes ao número de quartos e banheiros, número de lareiras, área total, sistemas de refrigeração e aquecimento, tipo de estacionamento, etc. Além disso, três variáveis referem-se às avaliações das escolas primárias, secundárias e de ensino médio da região, e outras oito referem-se à taxa de contrato de hipoteca, localização, data de construção e venda. O último atributo é a classe, que os autores converteram em binário: se o preço de fechamento é maior do que o preço de listagem ou vice-versa. Assim, o problema pode ser modelado como um problema de classificação a fim de decidir se um imóvel vale a pena de se investir. Para realizar essa classificação, os autores compararam diferentes algoritmos: árvores de decisão, RIPPER, Naive Bayes e AdaBoost. Os melhores resultados foram atingidos usando o RIPPER (*repeated incremental pruning to produce error reduction*) com um erro médio de cerca de 25

Ainda nesse ramo, Li *et al.* (2021) propuseram um *framework* de fusão de dados georreferenciados provenientes de várias fontes, integrando o *XGBoost* e o Modelo de Preços Hedônicos (HPM) a fim de estudar as relações complexas entre o preço dos imóveis e suas características individuais. Especificamente, ambos algoritmos foram capazes de revelar diferentes aspectos das relações complexas entre o preço entre os preços dos imóveis e os fatores de influência, classificando tais fatores por importância e determinando seus efeitos quantificados. Os resultados do *XGBoost* entregaram as cinco variáveis mais importantes para a precificação dos imóveis em Shenzhen, entre elas distância até o centro da cidade, índice de vista verde, densidade populacional, taxa de administração de propriedade e nível econômico.Tal trabalho demonstrou que o uso de *big data* no mercado urbano, aprendizado de máquina e métodos estatísticos espaciais são capazes de fornecer novas fontes de dados e possibilitam abordagens interdisciplinares para compreender melhor a distribuição de preços dos imóveis.

3 METODOLOGIA

Este capítulo aborda o enquadramento da pesquisa, material e métodos utilizados para o desenvolvimento do trabalho e os procedimentos metodológicos, bem como os seus detalhamentos. A Figura 8 apresenta um fluxograma detalhando o passo-a-passo dessa metodologia. São três etapas macro: coleta e preparação de dados, treinamento dos modelos e avaliação dos resultados. O código fonte usado para o processo de coleta, analise exploratória e treinamento e avaliação do modelo deste estudo estão disponíveis no repositório Github: Projeto Imobiliário

Figura 8 – Fluxograma com visão geral do projeto



Fonte: Autor (2024)

3.1 COLETA DE DADOS

Os dados deste trabalho foram obtidos por meio da extração de anúncios de venda e aluguel de imóveis no site VivaReal para a região de Florianópolis, de acordo com a quantidade de imóveis disponíveis no portal imobiliário para o mês de Janeiro

de 2024. Durante o processo de coleta de dados, não foi aplicada nenhuma ordenação no site para evitar qualquer viés na seleção dos imóveis com base em faixas de preços específicas. Essa abordagem foi adotada a fim de assegurar que o conjunto de dados fosse o mais representativo possível, abrangendo uma maior dispersão e variabilidade nos preços dos imóveis. No entanto, essa ação resultou em uma base de dados não homogênea, o que foi explorado, analisado e devidamente tratado ao longo do estudo.

Para a coleta de dados reais do mercado imobiliário de Florianópolis, elaborou-se um algoritmo de *web scraping*, utilizando a abordagem que segue os links de páginas da web e utilizando as ferramentas: *requests* e *BeautifulSoup*. A biblioteca *requests* é responsável por realizar o mapeamento do protocolo HTTP na semântica orientada a objetos do Python, enquanto a biblioteca *BeautifulSoup* auxilia na formatação da página HTML, fornecendo objetos Python facilmente iteráveis que representam estruturas XML. As informações são processadas e armazenadas em um banco de dados CSV para facilitar o acesso e a manipulação dos dados. Posteriormente, esses dados são analisados dentro do ambiente virtual do VSCode.

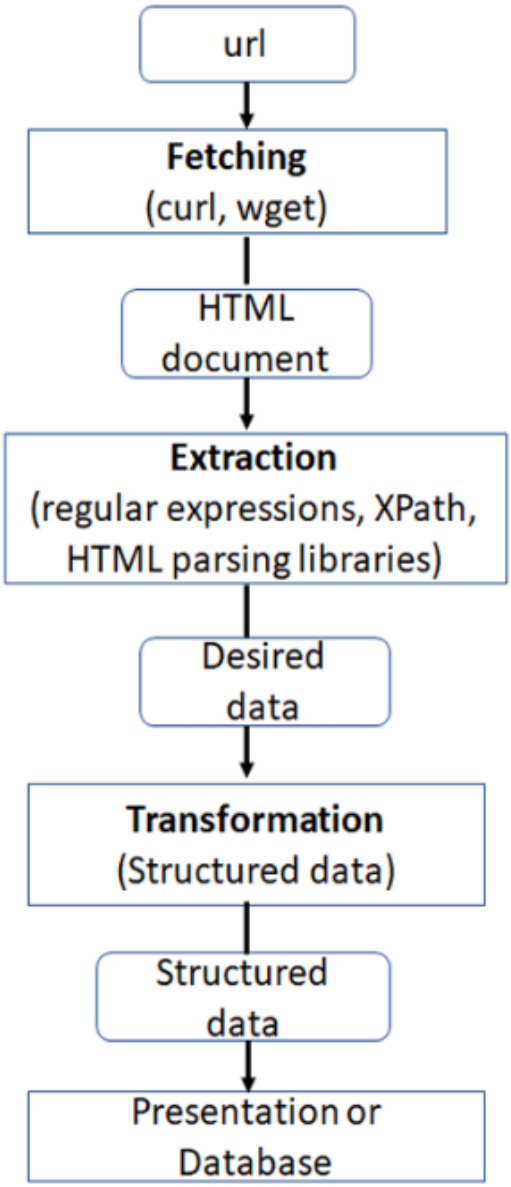
O trabalho será realizado no ambiente do *Jupyter Notebook*, documentos que contêm código executável, equações, visualizações e texto narrativo. Eles tornaram-se o sistema mais amplamente utilizado para programação literária interativa (SHEN, 2014).

3.1.1 Web Scraping

A pesquisa na web e a extração de informações geralmente são realizadas por *web crawlers*, o que segundo Kobayashi e Takeda (2000) é um programa ou script automatizado que navega na *World Wide Web* de maneira metódica e automatizada. O *Web Scraping*, segundo a definição de Khder (2021), é uma variação mais recente do *web crawler*, o qual lhe permite coletar dados de várias páginas *Web* em um formato estruturado para então ser posteriormente processado, analisado ou armazenado. A técnica de *web scraping* já está sendo utilizada em aplicações científicas e comerciais, uma vez que permite o desenvolvimento de banco de dados grandes e personalizados a custos mínimos. Ela tem como objetivo substituir métodos ultrapassados de coleta de dados, à medida que mais empresas buscam acompanhar as novas tendências seguidas pelos consumidores (HILLEN, 2019).

A parte prática da extração de dados é constituída de três etapas, conforme mostra a Figura 9.

Figura 9 – Processos do Web Scraping.



Fonte: Adaptado de (KHDER, 2021)

- Etapa de coleta: A fase de coleta consiste em acessar inicialmente o site desejado com as informações relevantes. Isso é realizado por meio do protocolo HTTP, que é um protocolo da Internet para enviar e receber solicitações de servidores da web. Navegadores da web utilizam métodos semelhantes para obter materiais em páginas da web. Nesta etapa, bibliotecas como curl2 e wget3 podem ser utilizadas para enviar uma solicitação HTTP GET para o endereço de destino (URL) e obter a página HTML como resposta (PERSSON, 2019).;
- Etapa de extração: Após coletar o HTML da página, extrai-se os dados de interesse. Nesta etapa, são utilizadas expressões regulares, bibliotecas de análise

de HTML e consultas XPath, que é uma linguagem especializada em localizar informações de documentos. Isso é conhecido como etapa de extração. O XPath é usado para buscar informações específicas em documentos. (PERSSON, 2019);

- Etapa de transformação: A última etapa consiste na conversão e estruturação dos dados de interesse em um formato estruturado para apresentação ou armazenamento em um banco de dados (PERSSON, 2019).

3.2 ANÁLISE EXPLORATÓRIA

Trata-se de uma etapa inicial essencial para compreender e resumir as principais características do conjunto de dados em estudo. Durante a análise exploratória, diversas técnicas estatísticas e visuais foram empregadas para identificar padrões, tendências, anomalias e possíveis relações entre as variáveis. Isso permitiu uma compreensão profunda dos dados antes da sua aplicação nos modelos de *machine learning*. A análise exploratória de dados foi conduzida para examinar a distribuição das variáveis, identificar valores ausentes ou discrepantes, explorar mais a fundo relações entre as variáveis e, em última análise, orientar as decisões sobre quais técnicas analíticas seriam mais apropriadas para atingir os objetivos da pesquisa.

3.3 PRÉ-PROCESSAMENTO DE DADOS

A fase de pré-processamento dos dados é fundamental para transformar os dados brutos em uma forma adequada para serem utilizados nos algoritmos de *machine learning*. Durante essa etapa, uma série de verificações de qualidade dos dados são conduzidas para assegurar sua integridade e consistência.

3.3.1 Remoção de duplicatas

No mercado imobiliário online, é comum encontrar o mesmo imóvel anunciado por diferentes corretores e imobiliárias em diversas plataformas. Isso resulta na ocorrência de múltiplos anúncios do mesmo imóvel dentro da mesma plataforma. Visando aprimorar a qualidade dos dados, os registros duplicados foram identificados e removidos do conjunto. Para isso, utilizou-se da função nativa do python *duplicated*, a qual verifica se cada linha é uma duplicata de qualquer linha anterior no conjunto de dados, levando em consideração todas as colunas.

3.3.2 Remoção de *outliers*

Para identificar os *outliers*, foram empregados histogramas e boxplots na escala logarítmica, a fim de visualizar e compreender a distribuição dos dados para cada variável quantitativa. Com base no conhecimento do domínio, foram estabelecidos

intervalos de valores considerados apropriados para o treinamento e validação do modelo. Para cada variável, definiu-se então um limite de corte, denominado *threshold*, de modo que todos os valores acima desse limiar fossem tratados como *outliers* e excluídos dos dados. Essa abordagem foi adotada visando garantir a integridade e a qualidade dos dados utilizados nas análises subsequentes.

3.3.3 Lidando com dados faltantes

Ao trabalharmos com modelos de *machine learning*, a qualidade dos dados desempenha um papel fundamental no desenvolvimento e na eficácia dos modelos preditivos. No entanto, é comum encontrar conjuntos de dados incompletos, nos quais algumas observações estão faltando valores para uma ou mais variáveis. Esses dados ausentes podem surgir devido a uma variedade de razões, como erros na coleta de dados, falhas no processo de medição ou campos deixados em branco durante a entrada de dados. Segundo Géron (2017), há duas diferentes técnicas para lidar com dados ausentes:

- Eliminar exemplos de treinamento ou características com valores ausentes: Um dos jeitos mais fáceis de se lidar com dados ausentes, no qual eliminamos a linha (ou coluna) cuja alguma observação esteja ausente. Entretanto, se removermos muitos dados, correremos o risco de perder informações valiosas que o modelo precise para discriminar os atributos.
- Imputação de valores ausentes: Caso a remoção de exemplos de treinamento ou a exclusão de colunas inteiras de características simplesmente não seja viável, é necessário recorrer a técnicas de interpolação para estimar os valores ausentes com base nos demais exemplos em nosso conjunto de dados. Uma das técnicas mais comuns é a imputação pela média ou pela mediana, na qual substituímos o valor ausente pela média/mediana de toda a coluna de características.

3.4 *FEATURE ENGINNERING*

Após realizado o pré-processamento dos dados, segue-se para a etapa de transformação dos dados (*Feature Enginnering*) com o intuito de gerar novos dados a partir de manipulações dos atributos já existentes. Por meio de técnicas como transformações matemáticas, criação de novas variáveis a partir de combinações de atributos existentes e extração de características relevantes, procurou-se enriquecer o conjunto de dados e capturar informações adicionais que pudessem ser úteis para auxiliar o modelo em suas predições.

3.4.1 Tratamento de variáveis categóricas

Nesta etapa, as variáveis categóricas foram tratadas por meio da transformação em variáveis *dummy*. Esse processo consiste em converter cada categoria de uma variável categórica em uma nova variável binária, onde 0 representa a ausência da categoria e 1 indica a presença. Tal abordagem é necessária, visto que os modelos de regressão só aceitam variáveis numéricas como entrada.

3.5 SEGMENTAR OS DADOS EM CONJUNTO DE TREINO E TESTE

Nessa etapa da metodologia, os dados foram segmentados em conjunto de treinamento e conjunto de teste. Segundo Raschka, Patterson e Nolet (2020), para evitar tanto o *overfitting* quanto o *underfitting*, a prática comum é dividir os dados em conjuntos de treinamento e teste, também chamada de *holdout*. O conjunto de treinamento é usado para ajustar o modelo aos dados disponíveis, enquanto o conjunto de teste é utilizado para avaliar o desempenho do modelo em dados não vistos anteriormente. ssa abordagem permite verificar se os modelos conseguem generalizar bem e fazer previsões precisas em novos conjuntos de dados, proporcionando uma avaliação mais confiável do desempenho dos modelos. Tal divisão pode ser feita de forma aleatória, sem considerar a distribuição das classes ou características dos dados, de forma estratificada, mantendo a proporção de classes entre os conjuntos.

3.6 TREINAMENTO DOS MODELOS

Após separar os dados em conjuntos de treinamento e validação, foram selecionados três modelos, denominados aqui como modelos candidatos, para aprender os padrões nos dados. Um modelo candidato representa um algoritmo ou configuração arquitetônica prospectiva considerada para resolver um problema específico. Optou-se por utilizar três modelos candidatos: XGBoost, Random Forest e Lasso Regression. Esses modelos foram escolhidos por representarem o estado da arte para tarefas de regressão em dados tabulares. A seleção de um modelo candidato visa encontrar um equilíbrio entre os requisitos do problema e os recursos computacionais e conhecimento disponíveis.

3.7 OTIMIZAÇÃO DE HIPERPARÂMETROS

Para a otimização dos hiperparâmetros, empregou-se o método de Random Search. O mesmo foi escolhido devido à sua capacidade de explorar eficientemente o espaço de hiperparâmetros. Em contraste com a busca em grade, que testa todas as combinações possíveis em uma grade predefinida, o Random Search seleciona aleatoriamente um conjunto limitado de combinações para avaliação. Essa abordagem

proporciona uma exploração mais eficiente, especialmente em espaços de hiperparâmetros de grande dimensão. Por meio de múltiplas iterações do Random Search, foram investigadas diversas combinações de hiperparâmetros de forma aleatória, buscando encontrar configurações que resultassem em melhor desempenho dos modelos de machine learning. Essa estratégia permitiu explorar o espaço de hiperparâmetros de maneira mais abrangente, identificando configurações ótimas ou próximas do ótimo em um tempo computacionalmente viável.

3.8 ANÁLISE DOS RESULTADOS

Depois de realizados os ajustes nas etapas anteriores, foi possível avançar para a última etapa do trabalho, que consiste na análise dos resultados. Durante essa etapa, todas as métricas obtidas nos modelos foram comparadas para selecionar aquele que apresentasse a melhor capacidade de previsão. As métricas utilizadas incluem RMSE, R^2 e MAPE. Além da comparação direta entre os modelos, também foram coletadas e analisadas as métricas entre os modelos que passaram por ajustes nos hiperparâmetros e aqueles que não passaram.

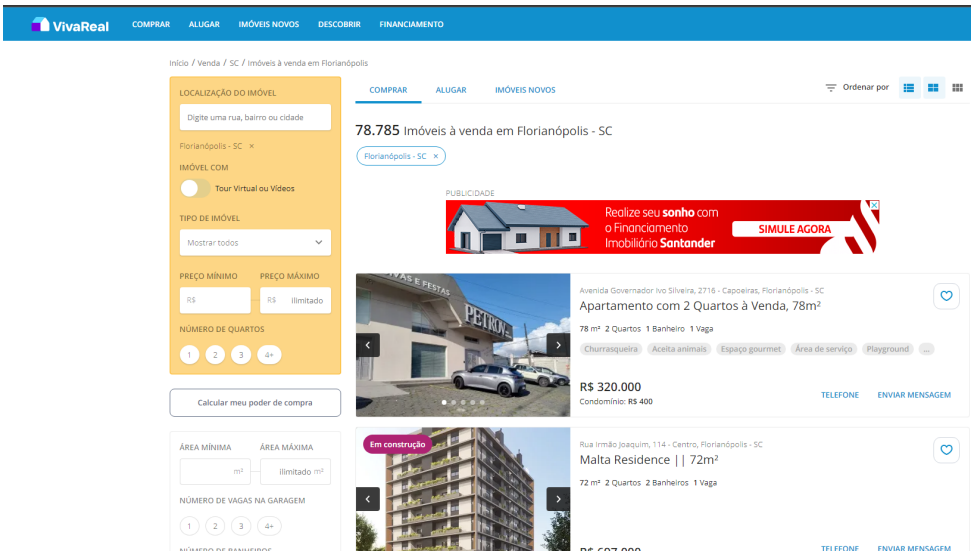
4 RESULTADOS

Neste capítulo, serão apresentados os resultados de cada etapa da metodologia. Todos os modelos de *machine learning* foram treinados utilizando o mesmo conjunto de dados para teste, treinamento e validação. Além disso, serão realizadas análises da importância de cada atributo para a variável preditora entre os modelos. Os resultados serão apresentados por meio de análises gráficas dos erros percentuais entre os preços previstos e os preços reais, juntamente com as métricas de desempenho, incluindo RMSE, R^2 e MAPE em forma tabular.

4.1 COLETA DE DADOS

A coleta de dados foi realizada por meio de técnicas de *web scraping* no site Viva Real, escolhido devido à sua estrutura mais permissiva aos algoritmos de extração de dados quando comparado a outros sites similares e por possuir uma ampla gama de atributos de para a análise (Figura 10).

Figura 10 – Portal imobiliário - Viva real



Fonte: Autor (2024)

Utilizando das principais bibliotecas em *Python* para *web scraping*, navegou-se por todas as páginas de anúncios de venda, filtrado para região de Florianópolis, coletando todas as informações disponíveis para cada anúncio registrado. Tais dados vieram no formato JSON, os quais foram agrupados e unificados em um único arquivo JSON, possibilitando sua manipulação e posterior análise. A tabela 1 mostra todas os atríbutos extraídos de cada anúncio no site.

Tabela 1 – Variáveis extraídas e formatos interpretados pelo Python

Coluna	**Valores não nulos**	**Tipo de variável**
displayAddressType	78960	object
amenities	78960	object
usableAreas	78960	object
constructionStatus	78960	object
listingType	78960	object
description	78960	object
title	78960	object
unitTypes	78960	object
nonActivationReason	78960	object
propertyType	78960	object
unitSubTypes	78960	object
id	78960	int64
portal	78960	object
parkingSpaces	78960	object
suites	78960	object
publicationType	78960	object
externalId	78960	object
bathrooms	78960	object
usageTypes	78960	object
totalAreas	78960	object
advertiserId	78960	object
bedrooms	78960	object
pricingInfos	78960	object
showPrice	78960	bool
status	78960	object
videoTourLink	78960	object
stamps	78960	object
address.country	78960	object
address.zipCode	78960	object
address.geoJson	78960	object
address.city	78960	object
address.streetNumber	46255	object
address.level	78960	object
address.precision	78960	object
address.confidence	78960	object

Continua na próxima página

Tabela 1 – Continuação

Coluna	Valores não nulos	Tipo de variável
address.stateAcronym	78960	object
address.source	78960	object
address.point.lon	46235	float64
address.point.source	46235	object
address.point.lat	46235	float64
address.ibgeCityId	78960	object
address.zone	78960	object
address.street	63161	object
address.locationId	78960	object
address.district	78960	object
address.name	78960	object
address.state	78960	object
address.neighborhood	78960	object
address.poisList	78960	object
address.pois	78960	object
address.valuableZones	78960	object

O banco de dados resultante da extração consistiu em 78.960 observações e 51 atributos, capturando 100% dos anúncios disponíveis no site naquele instante. A primeira etapa consistiu da remoção de características consideradas irrelevants para a predição do preço. Entre elas podemos listar diversas características compostas por apenas um valor, como constructionStatus igual a *None*, listingType igual a *USED*. Também foi removido atributos relacionados ao ID do anúncio, os quais tem sentido apenas para organização interna do sistema. Houve casos de colunas contendo apenas valores nulos, as quais também foram descartadas.

Após essa pré-limpeza, usando como critério apenas o significado de cada atributo, obteve-se um conjunto de dados de 78960 observações e 17 colunas, as quais serão posteriormente analisadas uma a uma para compreender sua relevância para a precificação, caso exista realmente essa relação. O conjunto do primeiro filtro de dados pode ser visto na tabela 2.

Tabela 2 – Variáveis escolhidas para análise exploratória

Coluna	Val. Não Nulos	Tipo de Variável	Descrição
amenities	78898	object	Listas com amenidades
usableAreas	78798	float64	Área útil em m²

Continua na próxima página

Tabela 2 – Continuação

Coluna	**Val. Não Nulos**	**Tipo de Variável**	**Descrição**
title	78898	object	Título do anúncio
parkingSpaces	71788	float64	Número de vagas de estacionamento
suites	67945	float64	Número de suítes
bathrooms	76151	float64	Número de banheiros
totalAreas	75553	float64	Área total do imóvel em m²
bedrooms	78898	object	Número de quartos
yearlyIptu	69455	float64	IPTU anual
price	78898	int64	Preço do imóvel
monthlyCondoFee	61598	float64	Taxa de condomínio mensal
address.zipCode	78960	object	CEP
address.streetNumber	46208	object	Número da rua
address.point.lon	46235	float64	Longitude
address.point.lat	46235	float64	Latitude
address.street	63108	object	Nome da rua
address.neighborhood	78898	object	Bairro

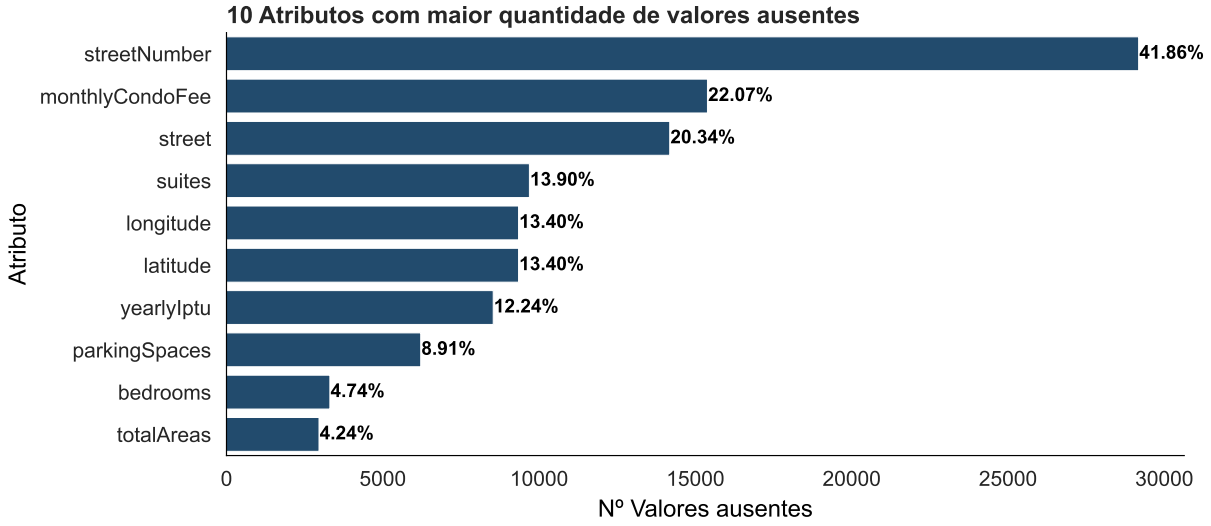
4.2 PRÉ-PROCESSAMENTO DOS DADOS

Neste capítulo, será realizada a análise exploratória dos dados, a qual incluirá a tratativa de valores ausentes, a identificação e o tratamento de outliers, além do processo de *feature engineering* para criação de novos atributos. A análise exploratória visa compreender a distribuição dos preços por metro quadrado em diferentes bairros, identificar padrões e anomalias, e gerar *insights* sobre o comportamento do mercado imobiliário em Florianópolis. A tratativa de valores ausentes garantirá a integridade da amostra, enquanto o tratamento de *outliers* minimiza distorções nos modelos preditivos. Por fim, o *feature engineering* aprimorará a representatividade das variáveis, potencializando a precisão dos modelos de *machine learning*.

Antes de começar a análise, a primeira etapa consistiu na remoção de duplicatas, visto que a presença de informações repetidas poderia introduzir um viés e distorcer os resultados do modelo. Encontrou-se 9081 observações de anúncio duplicado, os quais foram removidos, levando a uma redução de 11.51% do banco de dados original. Por fim, plotou-se as 10 variáveis com maior quantidade de valores nulos, afim de analisar se alguma variável seria impedida de entrar no modelo, conforme pode ser visto na figura 11. Apesar de 41% dos dados do atributo *streetNumber* e 20% de

street estarem ausentes, essas informações podem ser encontradas indiretamentes por meio do CEP do anúncio (disponível na coluna zipCode).

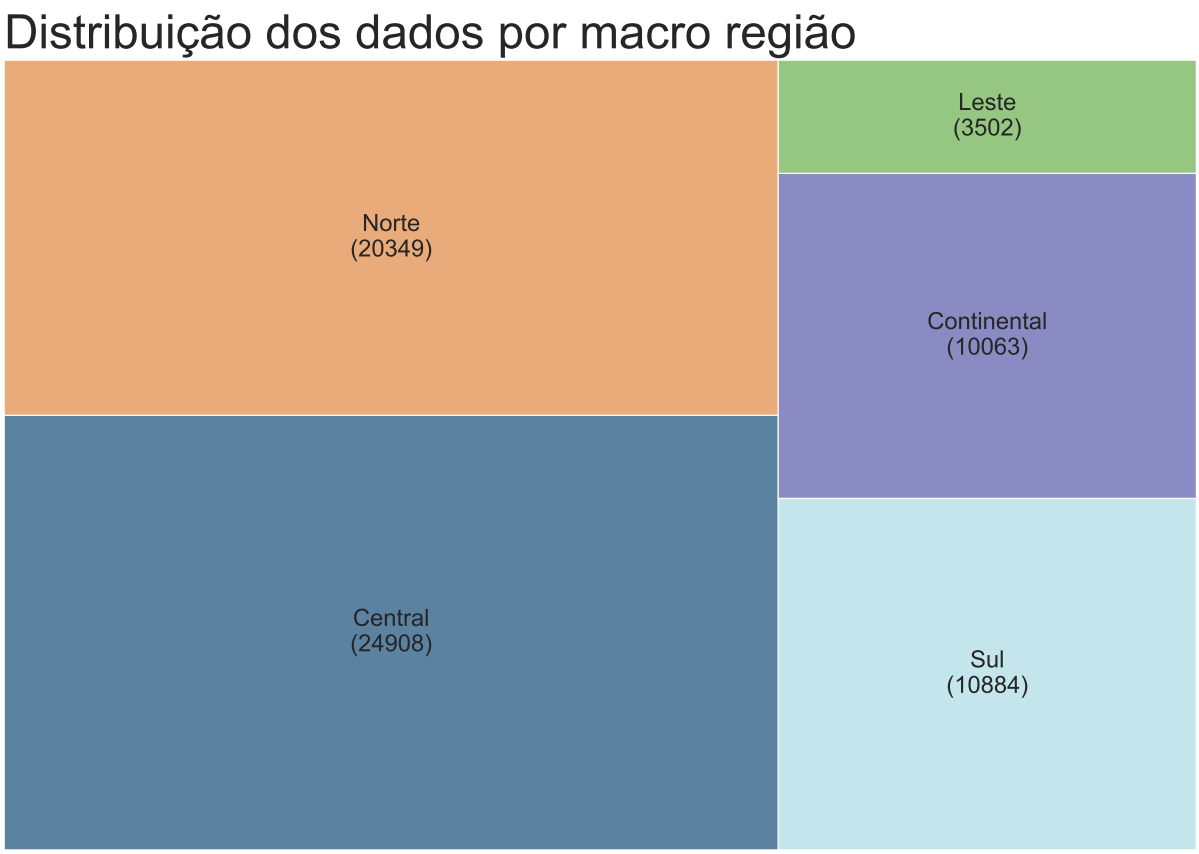
Figura 11 – Quantidade de valores nulos por atributo



Fonte: Autor (2024)

Outra análise importante para posterior entendimento do modelo foi na criação da coluna *macroRegions*, na qual foi agrupado cada bairro em sua devida macro região de Florianópolis. Pode-se observar pela figura 12 que a região central compõe 1/3 do banco de dados, enquanto a região Leste não tem muita representatividade.

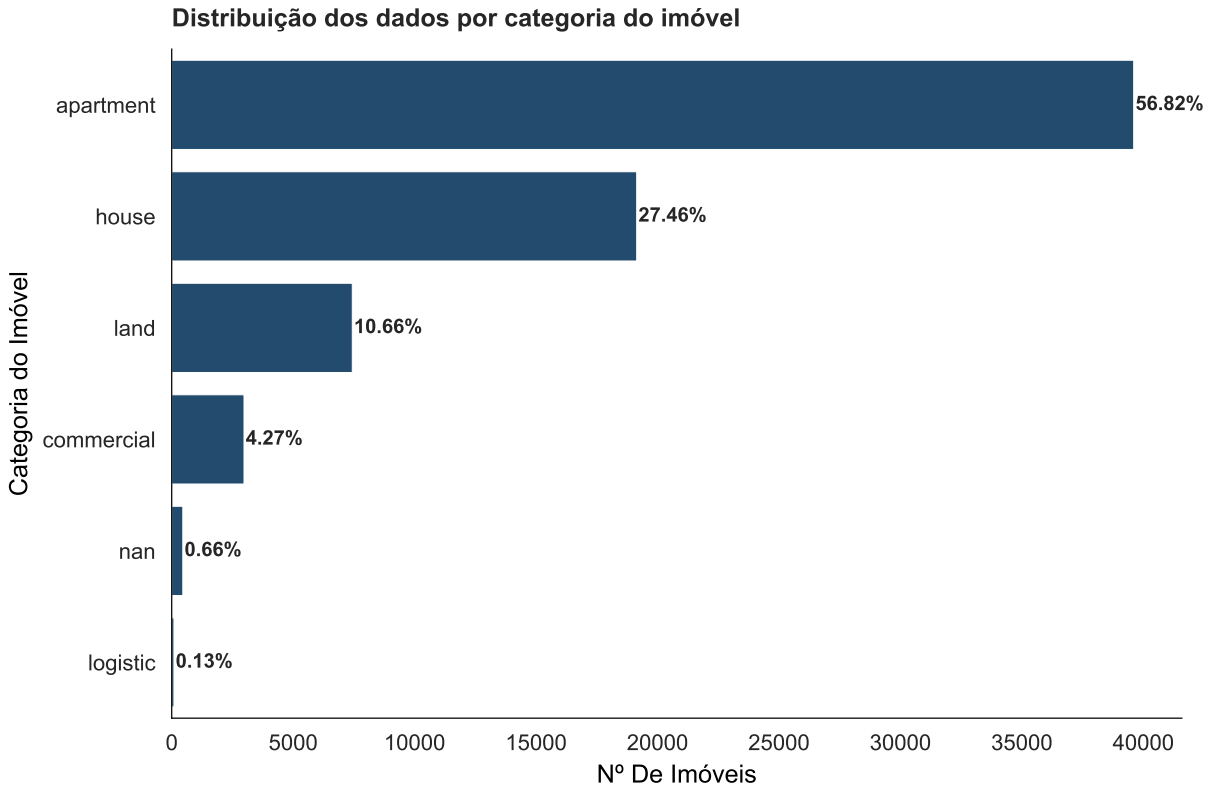
Figura 12 – Distribuição dos dados por macro região



Fonte: Autor (2024)

A partir da categoria 'unitTypesCategory', foi possível realizar outro agrupamento dos dados, identificando cinco principais tipos de imóveis (Figura 13): Apartamento, casa, terreno, comercial e logístico. Para o nosso conjunto de dados, observa-se que apartamento (39.604) e casa (19.143) totalizam 84.28% dos registros. A coluna NaN foi criada como uma variável *dummy* para agrupar as outras categorias que têm pouca representatividade no *dataframe* e não se encaixam em nenhuma das outras. Como o objetivo do modelo era a predição para bens de moradia, foram removidas as demais linhas que não correspondiam a apartamentos ou casas, resultando em um conjunto de dados reduzido para 58.747 registros (uma perda de 15.72%).

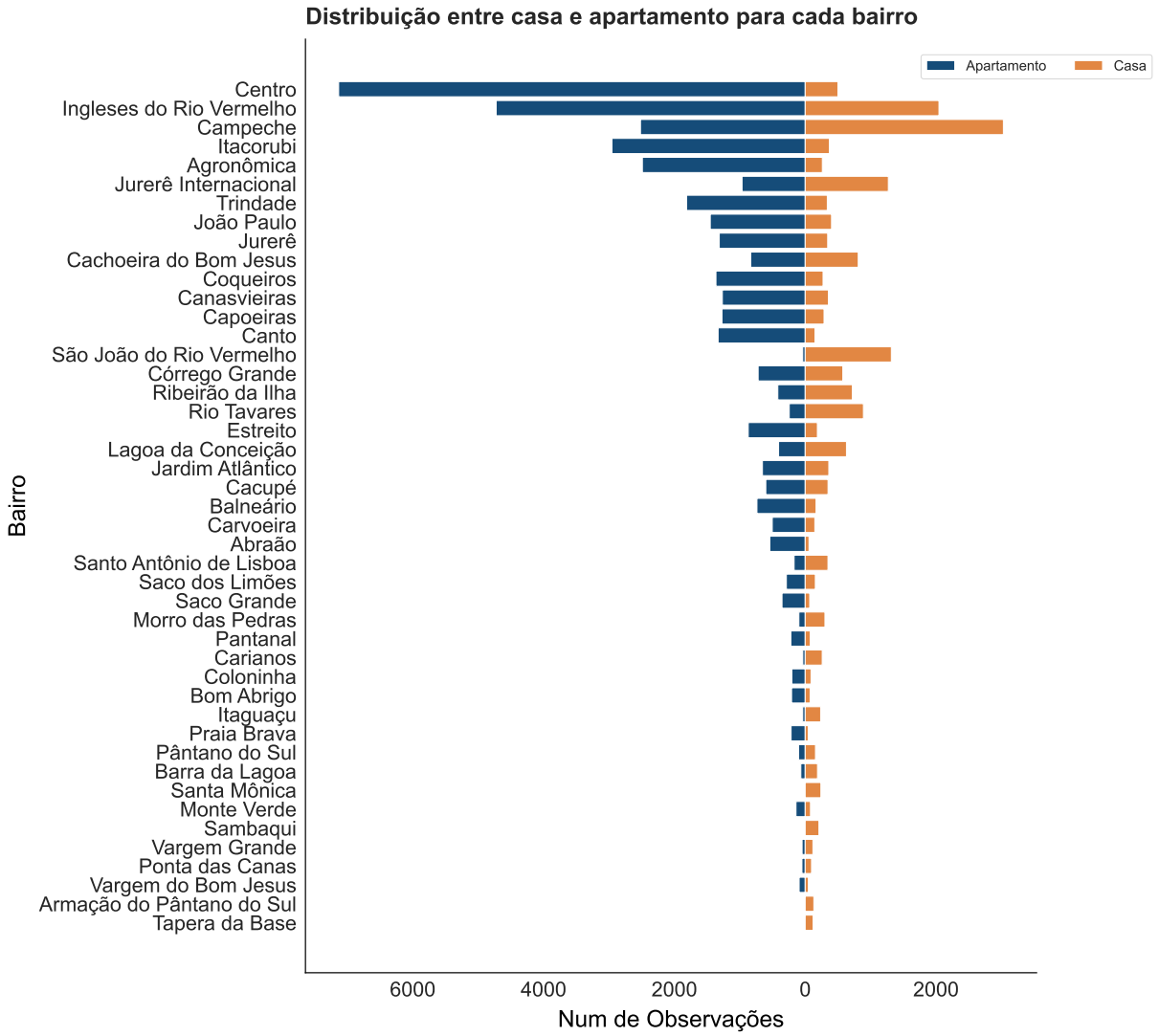
Figura 13 – Distribuição dos dados por tipo de anúncio



Fonte: Autor (2024)

Criou-se um gráfico de distribuição de dados para analisar a quantidade de observações por bairro. Essa visualização permite compreender melhor a qualidade da predição do algoritmo em cada bairro, dada sua proporção nos dados de treino e validação. Ao examinar a distribuição, é possível identificar desequilíbrios na quantidade de dados disponíveis por área geográfica. Tal etapa auxilia no entendimento de um possível *overfitting* do modelo, uma vez que bairros com um número significativamente menor de observações podem resultar em um ajuste excessivo do algoritmo para essas regiões específicas, levando a previsões enviesadas.

Figura 14 – Distribuição dos dados por bairro



Fonte: Autor (2024)

Conforme visto na figura 14, o Centro se mostra o bairro com maior representatividade do conjunto de dados. Apesar de alguns bairros como Tâpera, Armação e entre outros terem poucos anúncios, optou-se por não remover eles do conjunto de dados, afim de verificar no fim como o modelo genérico iria performar na precificação de imóveis destas regiões.

A próxima etapa da análise consistiu na exploração das variáveis numéricas relacionadas com o número de garagens, suítes, banheiros e quartos disponíveis nas propriedades. Para isto, utilizou-se o comando *'describe()'* do Python para obter uma descrição estatística resumida, conforme visto na tabela 4.

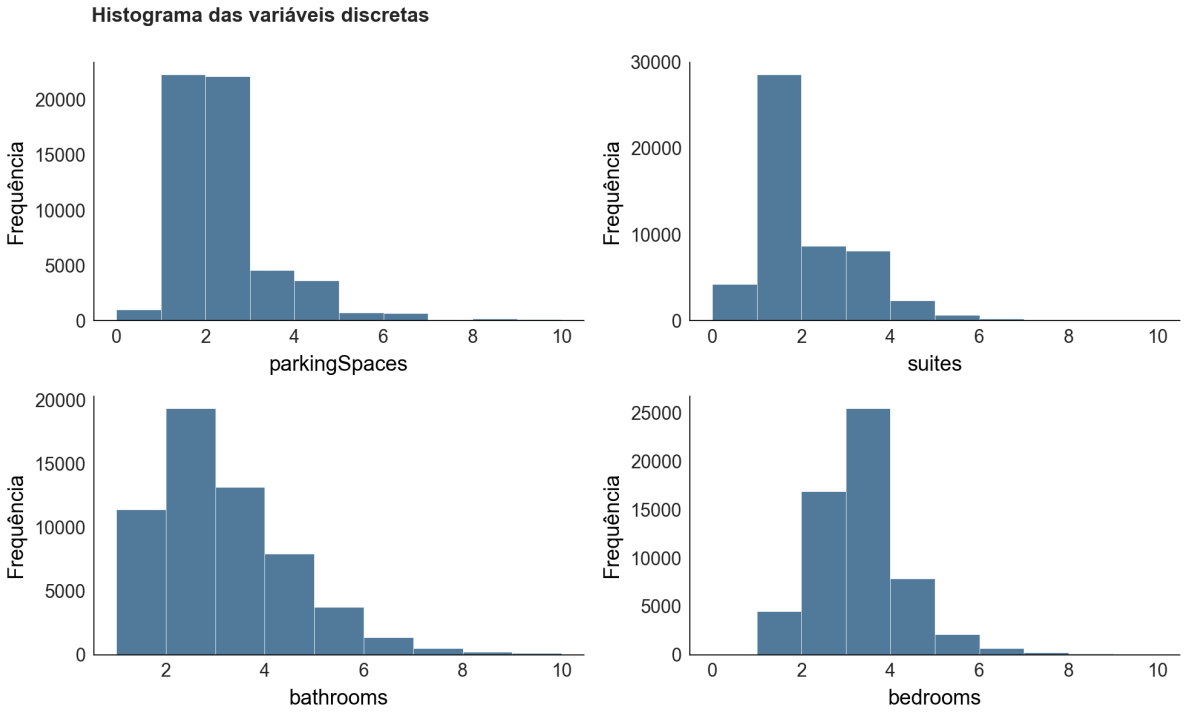
Tabela 3 – Análise descritiva: garagem, suites, banheiros e quartos

	parkingSpaces	**suites**	**bathrooms**	**bedrooms**
Count	56334	53620	58747	58681
Mean	2.422	1.625	2.725	2.868
Std	103.434	1.231	1.535	1.199
Min	0	0	1	0
25%	1	1	2	2
50%	2	1	2	3
75%	2	2	3	3
Max	24546	50	33	33

Durante a análise descritiva dos dados, observou-se que a variável garagens (*parkingSpaces*) apresentou uma média de 2.422 e um desvio padrão significativamente alto de 103.434. Essa disparidade foi acentuada pelo valor máximo de 24546, sendo um forte candidato a *outlier*. Uma investigação adicional revelou que esses valores extremamente altos estavam associados a anúncios como "Hotel para Venda"ou propriedades comerciais, os quais não se enquadram na categoria de imóveis residenciais e precisaram ser removidos. Além disso, foram encontradas outras anomalias, como anúncios de terrenos contendo valores de quartos e banheiros.

Para limpar os dados, a primeira etapa foi remover os anúncios que continham as palavras "terreno", "hotel"ou "comercial"em suas descrições, resultando na remoção de 420 observações. Em seguida, para a categoria "apartment", foram removidos os anúncios com mais de 6 quartos, banheiros, garagens, ou suítes, dada a distribuição de valores para este tipo de imóvel. Para casas, o critério de remoção foi definido como mais de 10 quartos, considerando que a maioria dos imóveis dessa categoria se encontrava nessa faixa de valores. Esta etapa resultou na remoção de 437 observações. A distribuição final destes atributos pode ser visto na figura 15.

Figura 15 – Histograma das variáveis discretas após filtragem.



Fonte: Autor (2024)

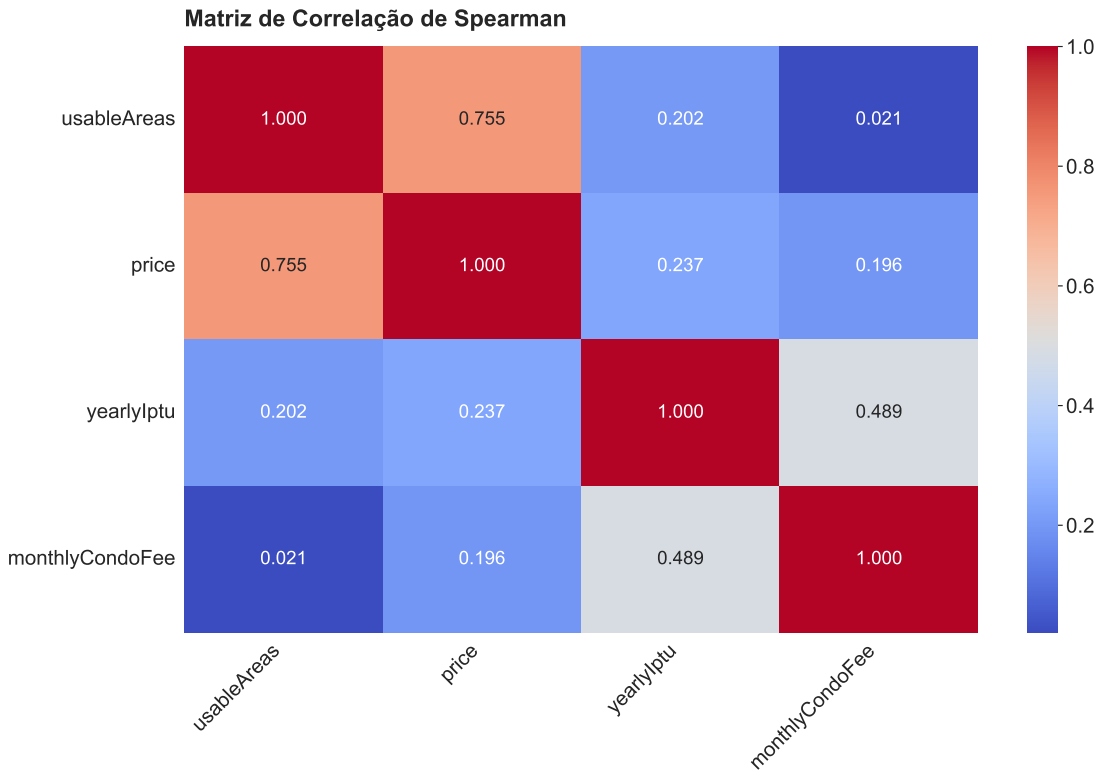
O mesmo processo foi repetido para as demais variáveis numéricas, como preço e área útil.

Tabela 4 – Análise descritiva: usableAreas, price, yearlyIptu e monthlyCondoFee

	usableAreas	**price**	**yearlyIptu**	**monthlyCondoFee**
Count	57.880	57.880	50.889	45.612
Mean	191	1.650.599	2859	2042
Std	2635	2.131.668	80.198	56.871
Min	10	500	0	0
25%	78	725.000	1	0
50%	117	1.150.000	650	530
75%	189	1.850.000	1524	900
Max	400.000	205.000.000	10.000.000	6.500.000

Para tratativa das variáveis continuas, o primeiro passo foi a remoção dos valores máximos, considerados aqui como *outliers*. Isso serviu para facilitar a visualização do histograma. Além disso, observa-se que o valor da média está muito diferente da mediana, indicando uma assimetria na distribuição das variáveis. Como tanto a variável *monthlyCondoFee* quanto *yearlyIptu* possuem diversos valores NaN em suas colunas, conforme visto anteriormente na figura 11, procurou-se entender a correlação dessas variáveis com o preço para determinar se seria melhor removê-las do conjunto de treino ou aplicar alguma técnica de imputação.

Figura 16 – Matriz de correlação de Spearmn



Fonte: Autor (2024)

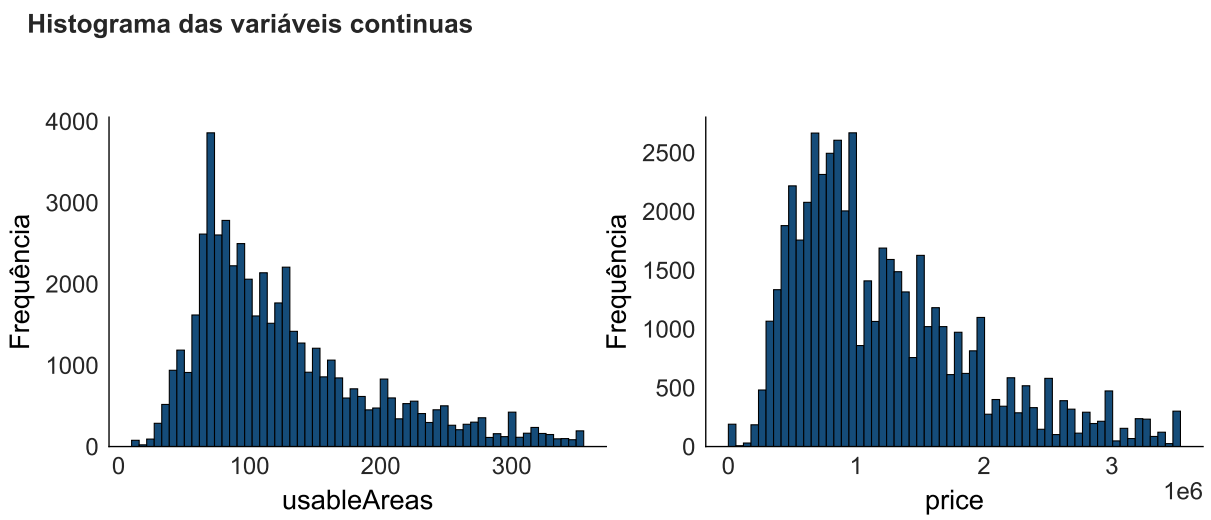
Utilizou-se da correlação de *Spearman* devido a presença de *outliers* nas variáveis, as quais prejudicam a perfomance de outros métodos como correlação de *Pearson*. Conforme visto na matriz 16, tanto o IPTU quanto condomínio possuem uma relação monotômica positiva, apesar de fraca. Isto indica que ambas não são preditores fortes para o modelo, pois embora haja uma tendência de preços mais altos em imóveis terem o IPTU mais alto, essa tendência não é forte o suficiente. Feita a remoção das duas variáveis, utilizou-se do intervalo interquatil (IQR) para remover imóveis cujo limite superior seja igual a $Q3 + 1.5 \times IQR$.

Tabela 5 – Remoção de outliers por IQR

	usableAreas	**price**
Limite Superior	355.5	3.537.500
Quantidade de Dados Removidos	3688	2244

A figura 17 mostra a distribuição final dos dados após a remoção dos *outliers*.

Figura 17 – Histograma para Preço e Área útil

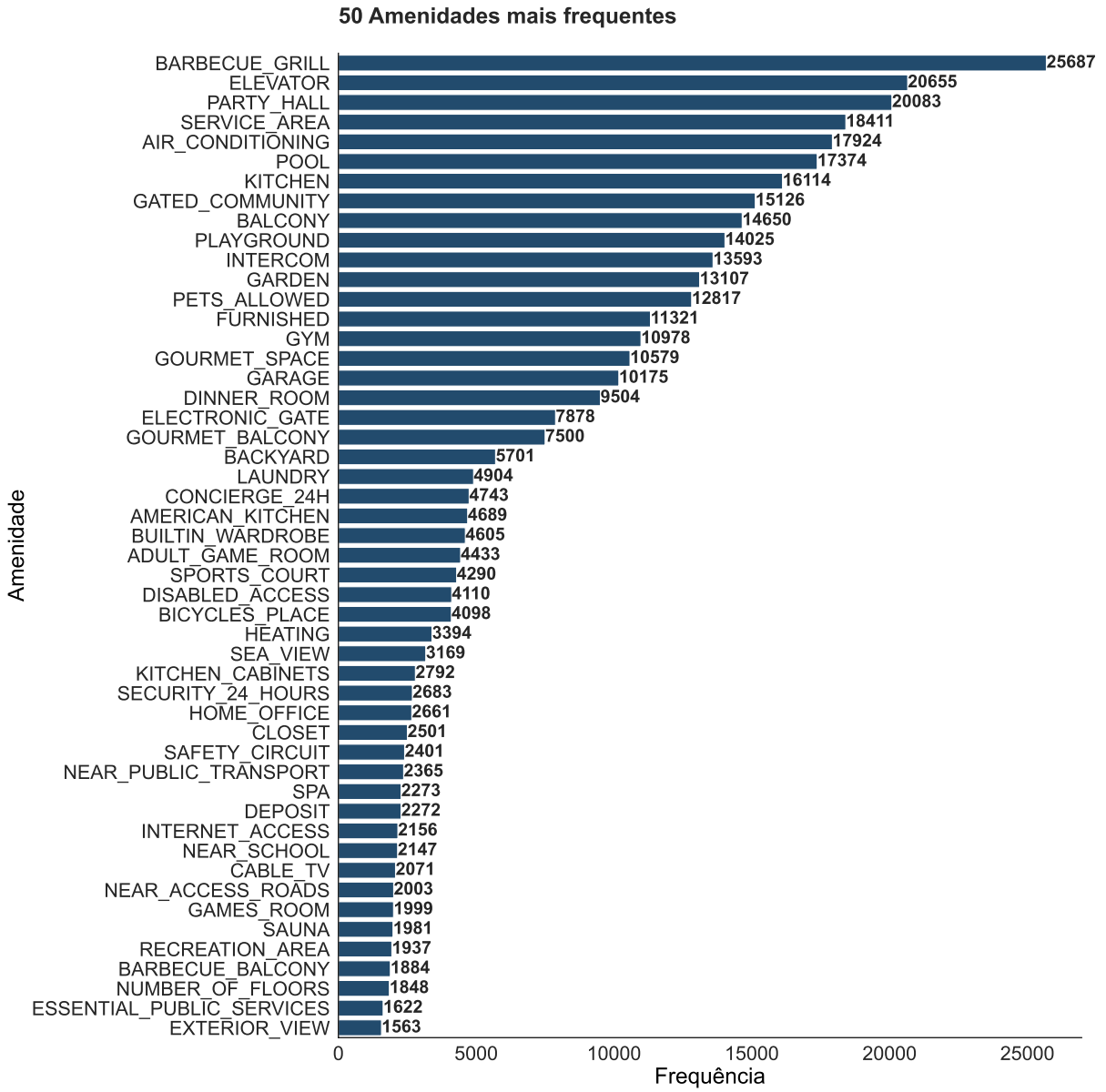


Fonte: Autor (2024)

Observa-se que a maior concentração de preços de venda está situada na faixa entre R$ 500.000,00 e R$ 1.500.000,00, com um pico de observações próximo a R$ 1.000.000,00. A assimetria à direita do histograma indica a presença de alguns valores atípicos com preços significativamente mais elevados, o que é comum no mercado imobiliário devido à existência de propriedades de luxo. Esse padrão de distribuição é consistente com a expectativa de que a maioria dos imóveis vendidos se situa em faixas de preço acessíveis a uma maior parte da população, enquanto uma parcela menor do mercado abrange imóveis de alto valor. Tal análise permite identificar a necessidade de estratégias específicas para tratar valores atípicos e segmentar o mercado de acordo com diferentes faixas de preço, o que pode melhorar a precisão das previsões dos modelos de aprendizagem de máquina empregados.

A próxima etapa consistiu na análise da variável *amenities*, a qual contém uma lista com as principais amenidades descritas para aquele imóvel. O gráfico da figura 18 mostra as 50 amenidades mais frequentes no conjunto de dados. As amenidades mais comuns incluem características como churrasqueira, elevador, salão de festas, área de serviço, ar condicionado, piscina entre outras.

Figura 18 – Número de observações para cada Amenidade



Fonte: Autor (2024)

Para incorporar essas informações nos modelos de *machine learning*, foi aplicado o processo de *feature engineering*. Esse processo transformou a coluna "amenities", que originalmente continha uma lista de palavras contendo as diferentes amenidades de cada imóvel, em variáveis binárias. Para cada uma das amenidades escolhidas, uma nova coluna foi criada no conjunto de dados. Essas colunas são representadas por valores binários, onde 0 indica a ausência da amenidade e 1 indica a sua presença no imóvel específico. As amenidades escolhidas para compor o conjunto de dados podem ser vistas na tabela 6

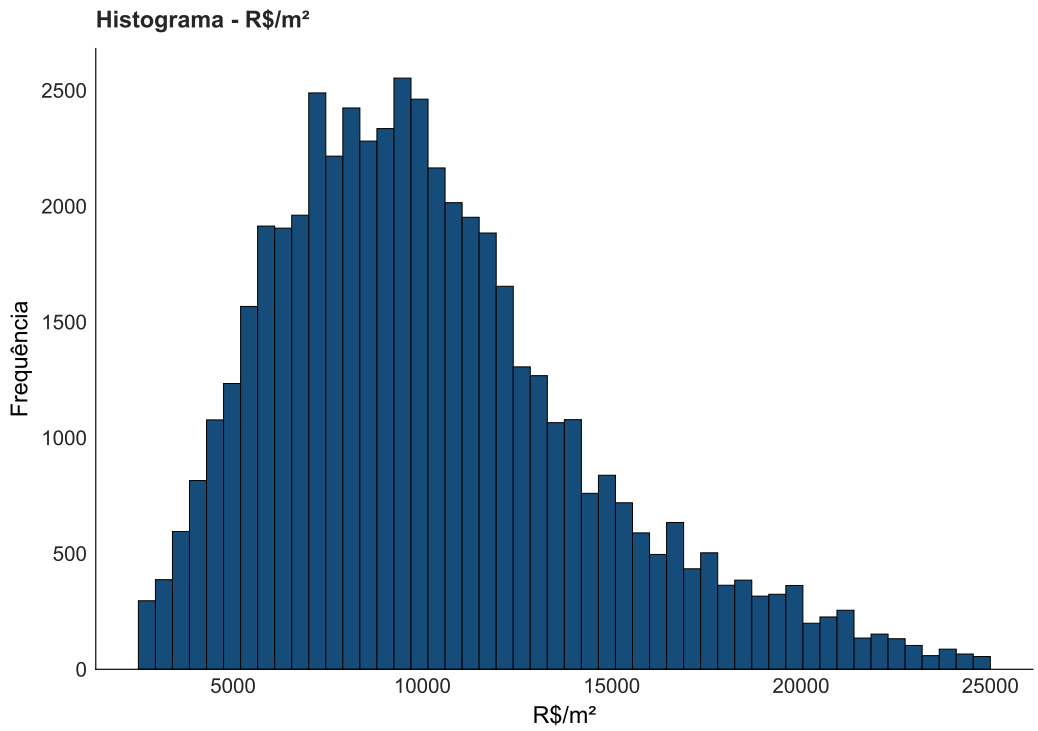
Tabela 6 – Tabela de Amenidades escolhidas

Amenidade	**Descrição**
GYM	Academia equipada disponível para os moradores.
BARBECUE GRILL	Churrasqueira para uso recreativo.
PARTY HALL	Salão de festas para eventos sociais.
POOL	Piscina para lazer e atividades aquáticas.
PLAYGROUND	Área de recreação para crianças.
GARDEN	Jardim ou área verde no condomínio.
GOURMET SPACE	Espaço gourmet para refeições e encontros sociais.
DINNER ROOM	Sala de jantar separada para refeições.
BACKYARD	Quintal privativo para uso exclusivo dos moradores.
SPORTS COURT	Quadra esportiva para práticas diversas.
SPA	Spa para relaxamento e cuidados pessoais.
GAMES ROOM	Sala de jogos com equipamentos diversos.
BARBECUE BALCONY	Varanda com churrasqueira.
RECREATION AREA	Área de lazer para atividades variadas.
SAUNA	Sauna para relaxamento.
GRASS	Área gramada para atividades ao ar livre.
CINEMA	Sala de cinema disponível para os moradores.
COWORKING	Espaço de coworking para trabalho remoto.
TENNIS COURT	Quadra de tênis para prática do esporte.
ELEVATOR	Elevador disponível no prédio.
CONCIERGE 24H	Serviço de portaria disponível 24 horas.
SECURITY 24 HOURS	Segurança disponível 24 horas por dia.

Além disso, a inclusão dessas características qualitativas busca validar a relevância das amenidades para a qualidade das predições do modelo. Ao transformar essas amenidades em variáveis binárias, é possível avaliar o impacto individual de cada uma delas no valor do imóvel. Isso permite identificar quais amenidades são mais significativas na determinação dos preços de venda. A análise da relevância dessas características pode revelar padrões e preferências dos compradores, auxiliando na tomada de decisões estratégicas por parte dos vendedores e financiadores, e proporcionando maior segurança e precisão nas avaliações imobiliárias.

Além de adicionar novas colunas ao banco de dados, também foram calculadas as métricas de preço por metro quadrado para cada observação. Isso foi feito dividindo os valores da coluna "Preço"pelos valores da coluna "Área". Logo em seguida, removeu-se todos os anúncios cujo preço por m² fosse menor que 2500 ou maior que 25.000, com o intuito de lidar com combinações de área e preço que fugissem do da distribuição. Seu histograma pode ser visto na figura 19.

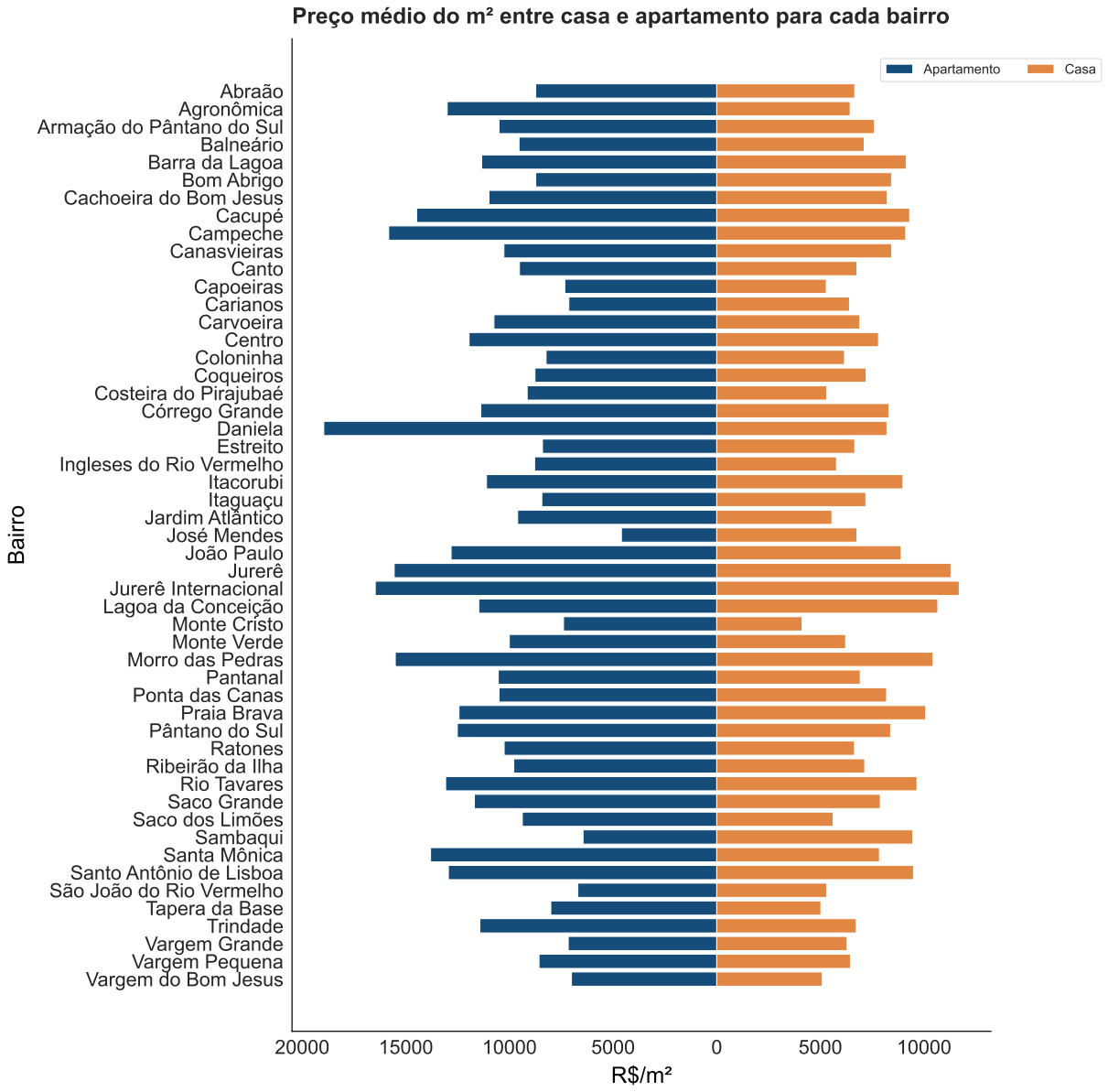
Figura 19 – Histograma - Preço por m²



Fonte: Autor (2024)

O gráfico da figura 20 compara os preços médios por metro quadrado entre casas e apartamentos em diversos bairros. A análise revela que, em geral, os apartamentos tendem a ter preços mais elevados em comparação às casas, com algumas exceções notáveis como Jurerê Internacional e Lagoa da Conceição, onde as casas alcançam valores muito altos, refletindo a exclusividade e o prestígio dessas localidades. Bairros como Centro e Agronômica também se destacam com preços elevados para apartamentos, o que pode estar relacionado à maior densidade populacional e à oferta de infraestrutura e serviços nessas áreas centrais.

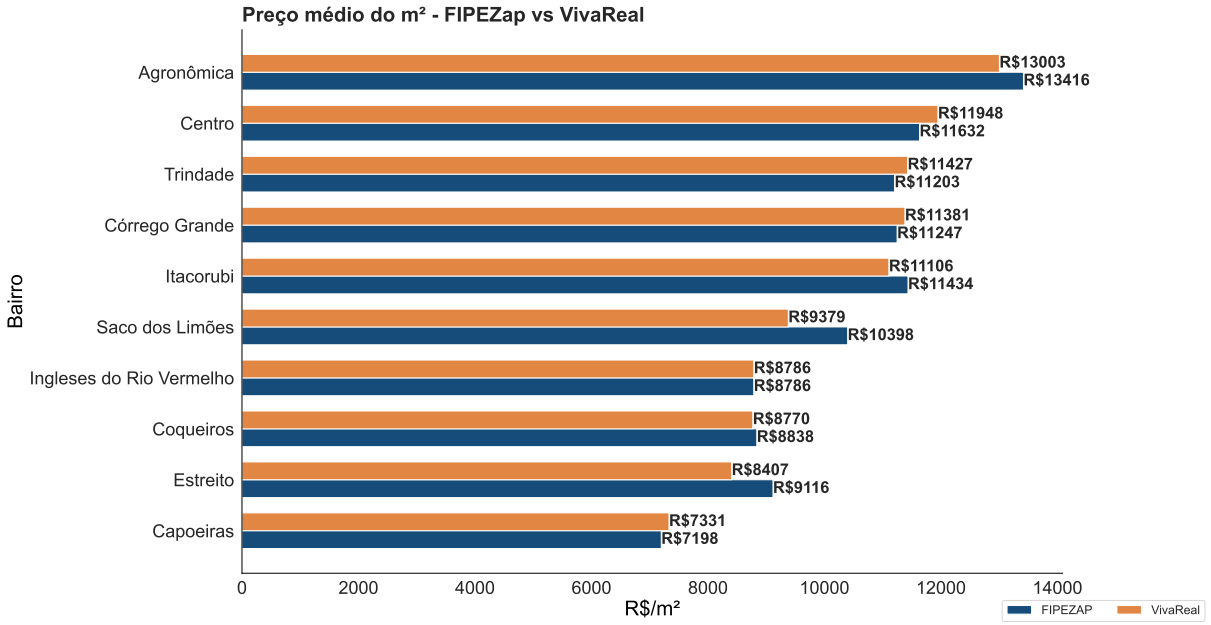
Figura 20 – Preço médio do m² entre casa e apartamento para cada bairro.



Fonte: Autor (2024)

Afim de validar o banco de dados obtido, utilizou-se as métricas de preço por m² para comparar com o índice FIPEZAP+ (FIPEZAP+ (SÃO PAULO)..., 2024). O mesmo é um indicador que acompanha o preço de venda de imóveis em várias cidades brasileiras. Ele é calculado pela Fundação Instituto de Pesquisas Econômicas (FIPE) em parceria com o portal de classificados ZAP.

Figura 21 – Comparativo do R$/m² entre o FipeZap e Viva Real.

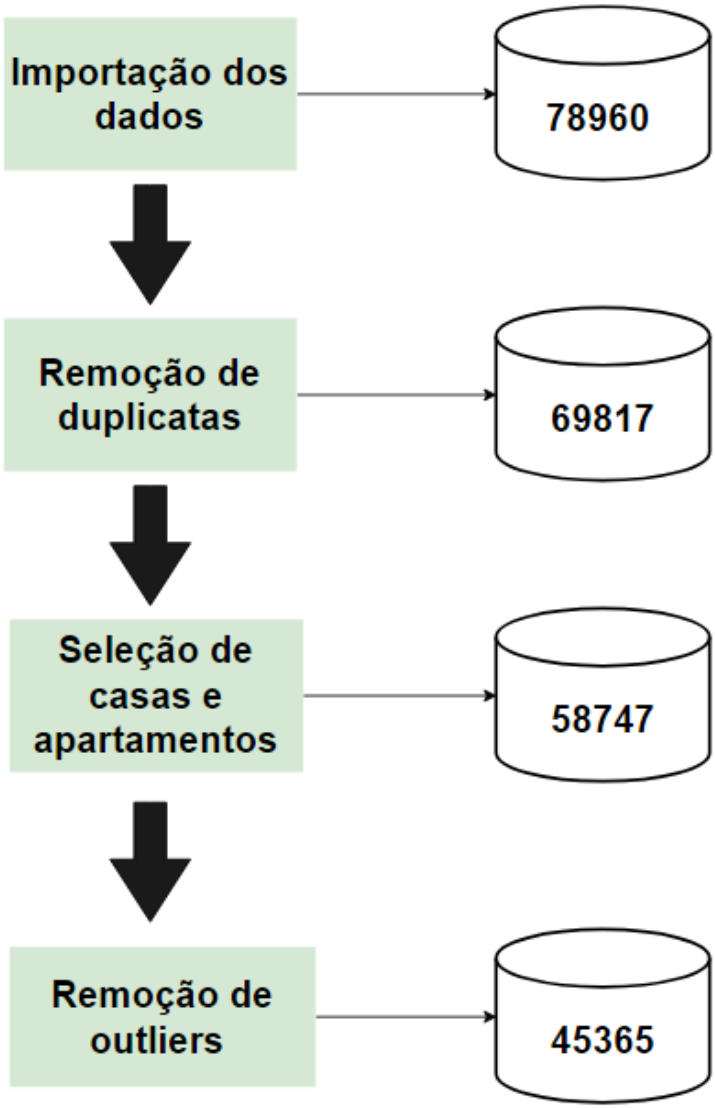


Fonte: Autor (2024)

O gráfico apresenta uma comparação dos preços médios por metro quadrado para os diferentes bairros de Florianópolis elencados pelo informe de março de 2024 dp FipeZAP. Como o relatório acompanha a variação do preço médio de apartamentos prontos em 50 cidades brasileiras, com base em anúncios veiculados na internet, utilizou-se para análise comparativa apenas o preço médio por m² para apartamentos. Observa-se que os valores entre as duas fontes de dados são bastante similares, o que confere uma segurança adicional sobre a qualidade e o tratamento correto dos dados no *dataset*. Bairros como Agronômica, Centro e Trindade mostram valores elevados, sugerindo uma alta demanda e uma maior valorização imobiliária nessas regiões. Em contraste, bairros como Capoeiras e Ingleses do Rio Vermelho apresentam preços mais baixos, indicando uma menor pressão de demanda ou diferentes características urbanísticas que afetam o valor imobiliário.

As principais etapas de remoção de dados e a quantidade de anúncios removidos em cada etapa pode ser vista resumidamente na figura 22. Após todas as etapas de exploração, limpeza, seleção de atributos e tratamento dos dados, obteve-se um conjunto de dados final com 45.365 observações e 29 atributos, os quais podem ser vistas na tabela 7.

Figura 22 – Etapas da limpeza de dados



Fonte: Autor (2024)

7 .

Tabela 7 – Descrição das variáveis

Coluna	Tipo de Dado	Descrição
usableAreas	Numérico	Área útil em metros quadrados
parkingSpaces	Numérico	Número de vagas de estacionamento
suites	Numérico	Número de suítes
bathrooms	Numérico	Número de banheiros
bedrooms	Numérico	Número de quartos
price	Numérico	Preço do imóvel

Continua na próxima página

Tabela 7 – Continuação

Coluna	**Tipo de Dado**	**Descrição**
neighborhood	Categórico	Bairro do imóvel
GYM	Binário	Academia equipada disponível para os moradores.
BARBECUE GRILL	Binário	Churrasqueira para uso recreativo.
PARTY HALL	Binário	Salão de festas para eventos sociais.
POOL	Binário	Piscina para lazer e atividades aquáticas.
PLAYGROUND	Binário	Área de recreação para crianças.
GARDEN	Binário	Jardim ou área verde no condomínio.
GOURMET SPACE	Binário	Espaço gourmet para refeições e encontros sociais.
DINNER ROOM	Binário	Sala de jantar separada para refeições.
BACKYARD	Binário	Quintal privativo para uso exclusivo dos moradores.
SPORTS COURT	Binário	Quadra esportiva para práticas diversas.
SPA	Binário	Spa para relaxamento e cuidados pessoais.
GAMES ROOM	Binário	Sala de jogos com equipamentos diversos.
BARBECUE BALCONY	Binário	Varanda com churrasqueira.
RECREATION AREA	Binário	Área de lazer para atividades variadas.
SAUNA	Binário	Sauna para relaxamento.
GRASS	Binário	Área gramada para atividades ao ar livre.
CINEMA	Binário	Sala de cinema disponível para os moradores.
COWORKING	Binário	Espaço de coworking para trabalho remoto.
TENNIS COURT	Binário	Quadra de tênis para prática do esporte.
ELEVATOR	Binário	Elevador disponível no prédio.
CONCIERGE 24H	Binário	Serviço de portaria disponível 24 horas.

Continua na próxima página

Tabela 7 – Continuação

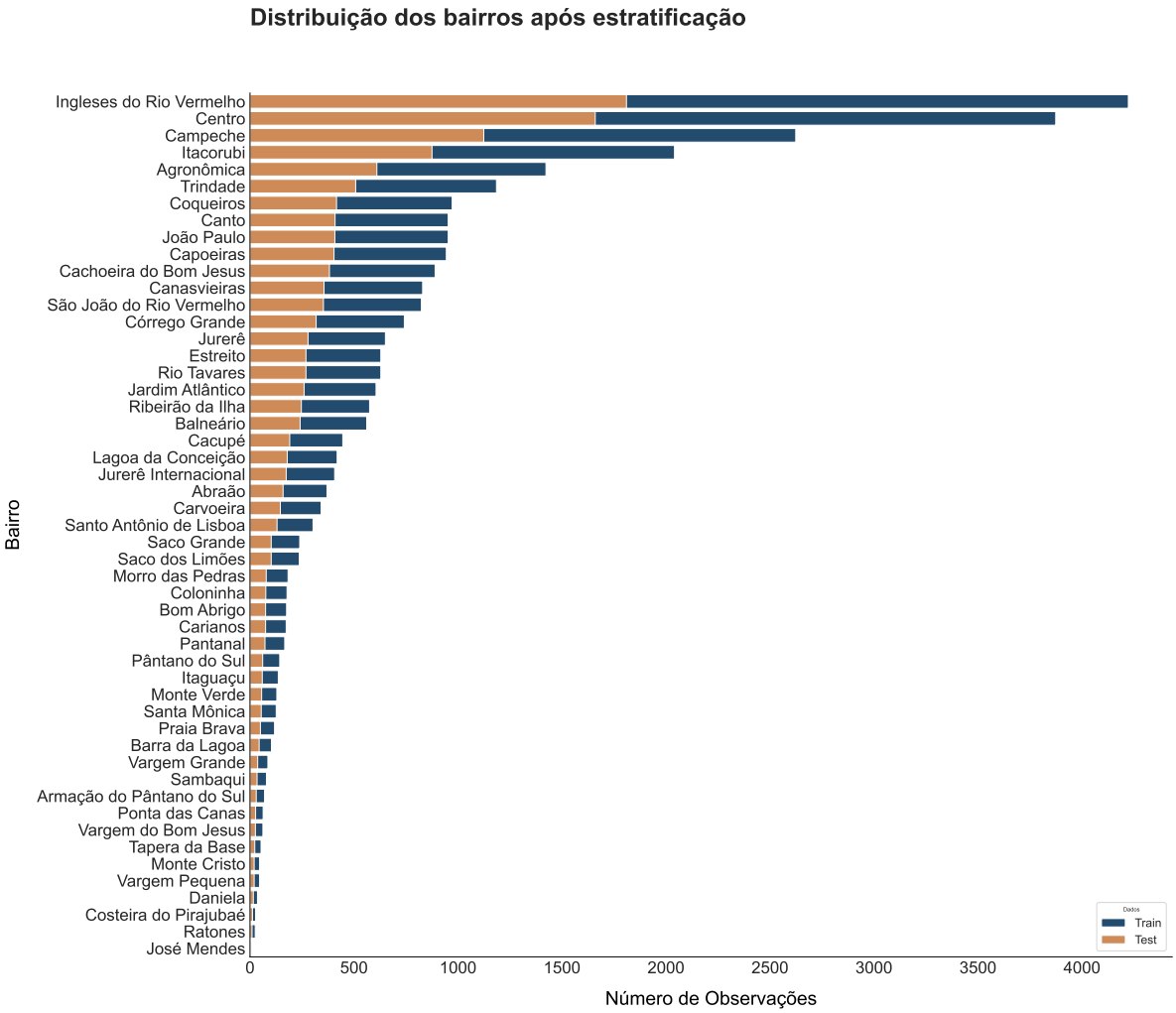
Coluna	**Tipo de Dado**	**Descrição**
SECURITY 24 HOURS	Binário	Segurança disponível 24 horas por dia.

4.3 SEPARAÇÃO DOS DADOS EM TREINO E TESTE

Para divisão dos dados em treino e teste, utilizou-se do método *holdout*, o qual pode ser resumido nas seguintes etapas. Primeiramente, o conjunto de dados foi separado em duas partes: um conjunto de treinamento e um conjunto de teste. Em seguida, um modelo foi ajustado aos dados de treinamento e as previsões foram feitas para o conjunto de teste. É essencial não treinar e avaliar um modelo no mesmo conjunto de treinamento, pois isso resulta em um problema de vazamento de dados. Esse vazamento ocorre quando informações dos dados de teste são inadvertidamente utilizadas durante o treinamento do modelo, levando a uma avaliação excessivamente otimista de desempenho. Como resultado, o modelo pode parecer ter uma capacidade de generalização melhor do que realmente possui quando confrontado com novos dados não vistos.

A divisão foi feita de forma estratificada por meio da coluna *neighboorhods*, o que significa que mantemos a mesma proporção de classes em ambas as divisões em relação aos bairros (Figura 23). Essa estratificação é importante para garantir que as distribuições das classes sejam preservadas tanto no conjunto de treinamento quanto no conjunto de teste. Definiu-se uma proporção de 70% dos dados para treino e 30% para teste.

Figura 23 – Separação dos dados em treino e teste



Fonte: Autor (2024)

4.4 OTIMIZAÇÃO DE HIPERPARÂMETROS

Na etapa de otimização de hiperparâmetros, foi utilizado o método *random search*, que é uma técnica de busca aleatória que explora de forma mais ágil o espaço de hiperparâmetros para um modelo. O *random search* funciona selecionando aleatoriamente combinações de valores para os hiperparâmetros especificados e avaliando o desempenho do modelo para cada conjunto de hiperparâmetros. Diferentemente de métodos de busca exaustiva, como o *grid search*, o *random search* não testa todas as combinações possíveis, o que o torna mais eficiente computacionalmente, especialmente em espaços de hiperparâmetros de alta dimensionalidade, a custo de não necessariamente encontrar a melhor combinação para aquele intervalo de valores. Nesta seção são descritos o signficado de cada hiperparâmetro, assim como o intervalo de valores utilizados e o hiperparâmetro que fornece o melhor resultado para o modelo.

Como principais hiperparâmetros para o *Random Forest* temos:

- ***n_estimators***: Número de árvores usadas no modelo.
- ***min_samples_split***: Número mínimo de amostras necessárias para dividir um nó interno.
- ***min_samples_leaf***: Quantidade mínima de amostras necessárias para serem consideradas como folha (não subdivididas).
- ***max_features***: Número máximo de features a serem consideradas ao procurar a melhor divisão.
- ***max_depth***: Profundidade máxima da árvore.
- ***bootstrap***: Indica se as amostras são amostradas com substituição (True) ou não (False) ao construir árvores.

Tabela 8 – Valores testados para hiperparâmetros - Random Forest

Hiperparâmetro	**Intervalo de Valores**	**Melhor resultado**
n_estimators	[200, 500]	273
min_samples_split	[2, 5, 10]	5
min_samples_leaf	[1, 2, 4]	1
max_features	[5, 7, 9]	9
max_depth	[2,18]	18
bootstrap	Booleano	False

Fonte: Elaborado pelo autor.

Como principais hiperparâmetros para o *XGboost* temos:

- ***objectives***: Define o tipo de objetivo de regressão a ser otimizado durante o treinamento.
- ***n_estimators***: Número de árvores a serem usadas no modelo.
- ***max_depth***: Profundidade máxima da árvore. Árvores mais profundas podem capturar relações mais complexas nos dados, mas também têm maior probabilidade de overfitting.
- ***min_child_weight***: Define a soma mínima do peso das amostras necessárias em um nó para continuar a dividir-se. Valores maiores podem ajudar a evitar overfitting.
- ***learning_rate***: Taxa de aprendizado, que controla a magnitude das atualizações de modelo durante o treinamento.

- ***subsample***: Proporção das amostras usadas para treinar cada árvore. Valores menores introduzem aleatoriedade e podem ajudar a reduzir o overfitting.
- ***colsample_bytree***: Proporção de features usadas para treinar cada árvore.
- ***gamma***: Parâmetro de regularização que controla a redução da profundidade da árvore. Valores maiores levam a uma poda mais agressiva das árvores.
- ***reg_alpha***: Termo de regularização L1 nos pesos das features. Ajuda a evitar overfitting, penalizando pesos maiores.
- ***reg_lambda***: Termo de regularização L2 nos pesos das features. Funciona de forma semelhante ao *reg_alpha*, mas penalizando pesos quadráticos maiores.

Tabela 9 – Valores testados para hiperparâmetros - XGBoost

Hiperparâmetro	**Intervalo de Valores**	**Melhor resultado**
objectives	reg:linear	reg:linear
n_estimators	[200, 500]	248
max_depth	[2,18]	18
min_child_weight	[2, 3, 4]	4
learning_rate	[0.045, 0.05, 0.06]	0.045
subsample	[0.5, 0.55, 0.85]	0.55
colsample_bytree	[0.6, 0.8, 1.0]	0.6
gamma	[0, 0.001, 0.01, 0.1, 0.2]	0.01
reg_alpha	[0, 0.5, 1]	1
reg_lambda	[0, 0.5, 1]	1

Fonte: Elaborado pelo autor.

Como principais hiperparâmetros para a *Lasso Regression* temos:

- ***alpha***: Parâmetro de regularização, que controla a força da penalização L1 (também conhecida como regularização Lasso). Valores maiores de alpha resultam em mais coeficientes sendo reduzidos exatamente para zero, o que leva a um modelo mais simples.
- ***selection***: Método usado para selecionar entre os coeficientes quando a atualização é feita. O método "cyclic"seleciona os coeficientes em uma ordem cíclica, enquanto o método "random"os seleciona aleatoriamente a cada vez.

Tabela 10 – Valores testados para hiperparâmetros - Lasso Regression

Hiperparâmetro	**Intervalo de Valores**	**Melhor resultado**
alpha	[0.1, 1]	0.94
selection	["cyclic", "random"]	"random"

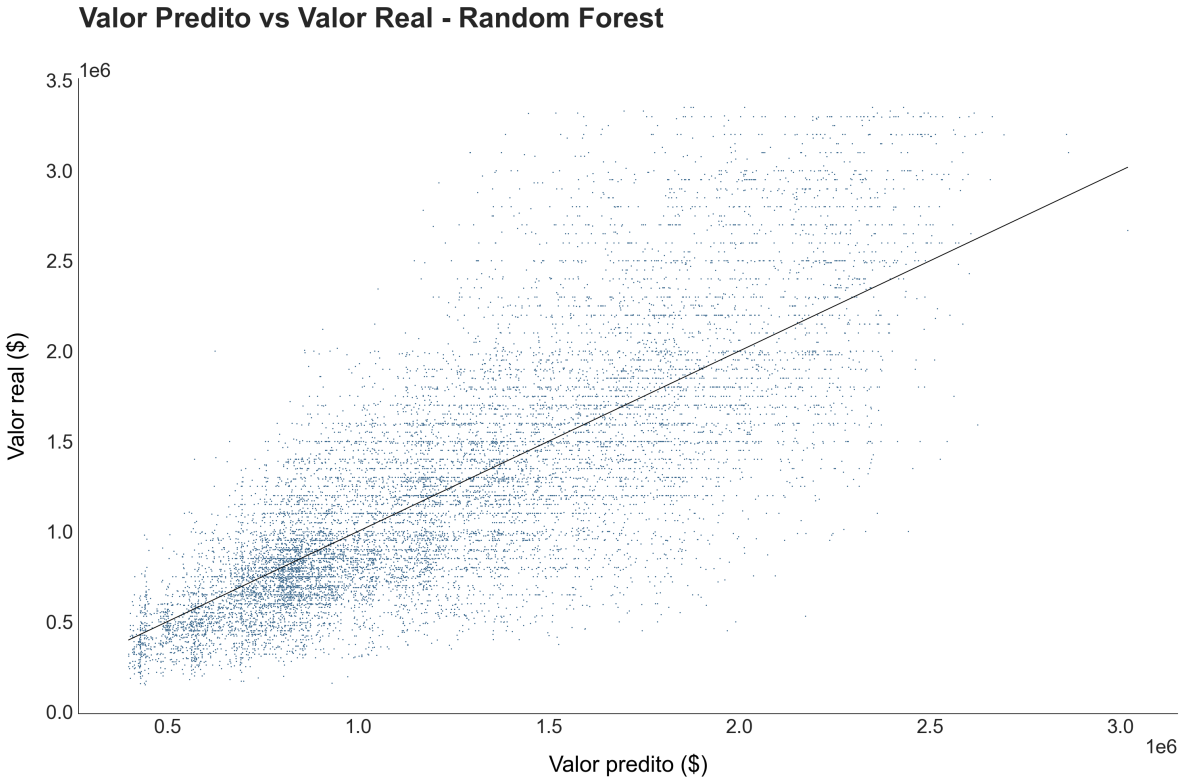
Fonte: Elaborado pelo autor.

4.5 ANÁLISE DOS RESULTADOS

Os algoritmos utilizados no estudo foram treinados e testados com o uso de hiperparâmetros pré-definidos e otimizados. Essa metodologia permitiu uma comparação entre os algoritmos, visando determinar qual deles apresentava o melhor desempenho na previsão de preços de imóveis para o conjunto de dados. Para essa avaliação, foram empregadas métricas como RMSE (Root Mean Square Error), R^2 (Coeficiente de Determinação) e MAPE (Mean Absolute Percentage Error).

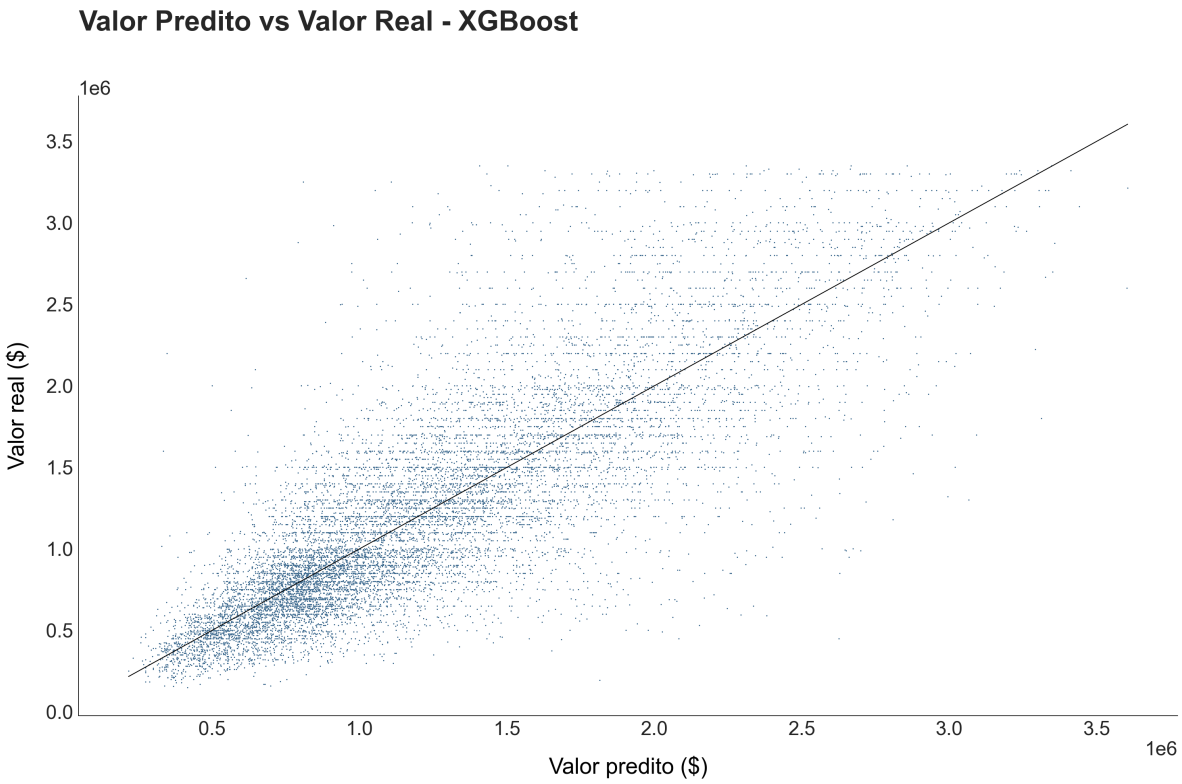
Para cada modelo avaliado tunado, gerou-se um gráfico de dispersão com o objetivo de visualizar a diferença entre o valor previsto e o valor real dos imóveis. Esses gráficos proporcionaram uma análise visual direta do desempenho dos modelos, permitindo identificar padrões de sub ou superestimação, bem como a dispersão dos pontos ao redor da linha de referência ideal, onde os valores previstos são iguais aos valores reais (Figuras 24, 25, 26)

Figura 24 – Gráfico de dispersão entre valores reais e valores previstos - Random Forest



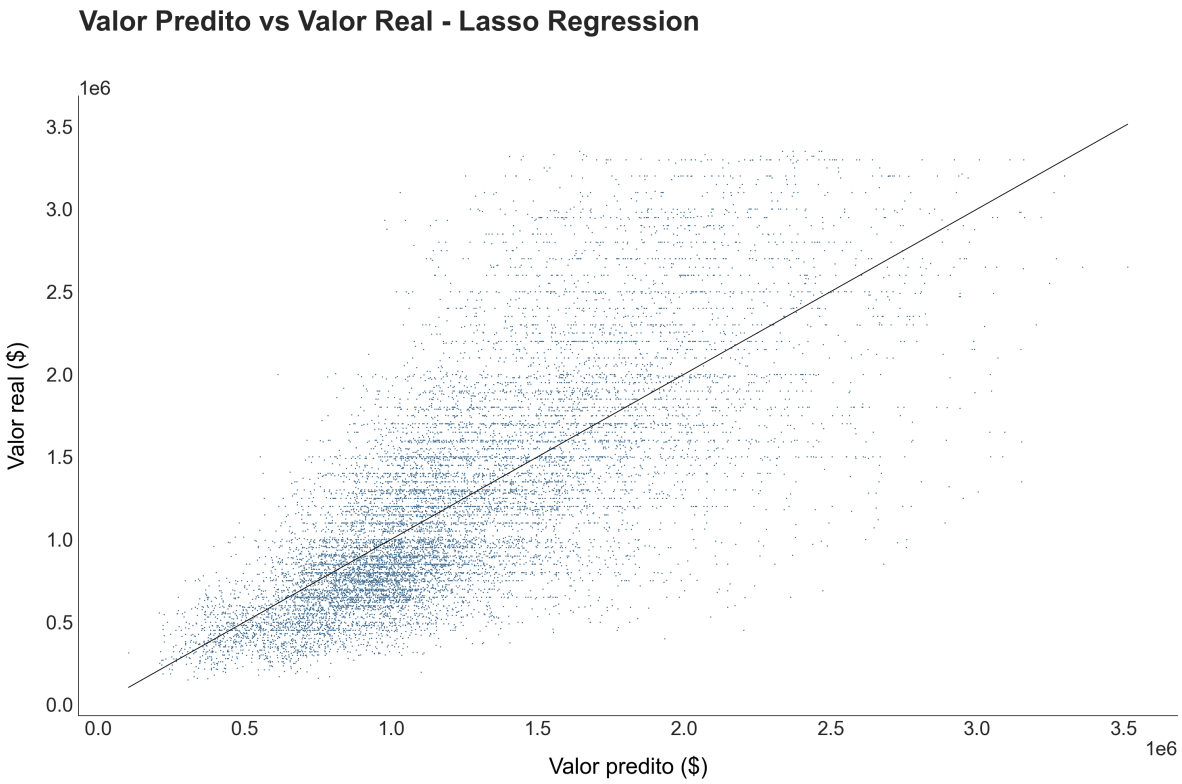
Fonte: Autor (2024)

Figura 25 – Gráfico de dispersão entre valores reais e valores previstos - XGBoost



Fonte: Autor (2024)

Figura 26 – Gráfico de dispersão entre valores reais e valores previstos - Lasso



Fonte: Autor (2024)

A análise comparativa dos gráficos de dispersão dos modelos Random Forest, XGBoost e Lasso Regression indica que todos os modelos apresentam boa precisão para valores preditos mais baixos, mas variam em desempenho para valores mais elevados. O modelo XGBoost demonstrou uma leve vantagem com menor dispersão para valores altos, seguido pelo Random Forest. O modelo Lasso Regression apresentou maior variabilidade e erros de predição, especialmente para valores mais altos, indicando limitações significativas neste contexto. A tabela 11 resume as métricas mencionadas para os modelos com parâmetros genéricos e otimizados, exibindo os melhores desempenhos obtidos.

Tabela 11 – Resumo das métricas para os modelos

Modelo	R^2	RMSE	MAPE
Random Forest	0.62	422846	29.42%
Random Forest (Tunado)	0.69	370975	24.93%
XGBoost	0.67	394139	25.17%
XGBoost (Tunado)	0.70	367513	24.12%
Lasso Regression	0.54	463778	32.20%
Lasso Regression (Tunado)	0.56	440134	31.62%

O modelo XGBoost com hiperparâmetros ajustados obteve o melhor desempe-

nho, com um R^2 de 0.70, indicando que cerca de 70% da variabilidade nos preços dos imóveis foi explicada pelo modelo. Além disso, o erro médio quadrático (RMSE) para esse modelo sugere que as previsões estão desviadas em cerca de R$367.513 em relação aos preços reais dos imóveis. Por fim, o erro percentual absoluto médio (MAPE) para o XGBoost ajustado foi de 24.12%, indicando que as previsões do modelo estão, em média, desviadas cerca de 24.12% em relação aos preços reais dos imóveis.

Comparando com o estudo de Magri Zaghi *et al.* (2023) para um conjunto de dados similar, o modelo Random forest ajustado apresentou o melhor desempenho para métricas usadas, como um R^2 de 0.818, RMSE (R$516.000) e MAPE (25,84%), estando levemente a frente do XGBoost. Já Xu e Nguyen (2022) teve o XGBoost como modelo mais eficaz na precificação de imóveis, dentre os 5 diferentes treinados, estando novamente próximo do Random forest. Tais resultados mostram que ambos os modelos são particularmente adequados para problemas de precificação de imóveis pois conseguem lidar bem com a heterogeneidade dos dados imobiliários, que frequentemente incluem variáveis categóricas e numéricas, além de interações complexas entre essas variáveis.

Logo em seguida, aplicou-se as métricas aos dados de teste para os bairros segmentados, com o objetivo de analisar o comportamento do modelo ajustado para cada um deles. Essa análise granular permitiu uma compreensão mais detalhada do desempenho dos algoritmos em diferentes contextos geográficos, revelando possíveis variações na capacidade preditiva dos modelos em relação às características específicas de cada bairro, conforme pode ser visto na tabela 12.

Tabela 12 – Resumo das métricas por bairro ordenado por R^2.

Bairro	**R^2**	**RMSE**	**MAPE**	**Observações**
Itacorubi	0.81	252900.82	14.97	874
Córrego Grande	0.80	303426.81	15.58	313
Coqueiros	0.79	285738.28	21.62	410
Trindade	0.78	262781.63	18.52	506
Carvoeira	0.76	232751.85	21.05	146
João Paulo	0.75	367041.06	17.04	430
Barra da Lagoa	0.72	454786.41	21.58	43
Balneário	0.72	328323.52	22.25	240
Centro	0.71	398559.00	18.85	1756
Santo Antônio de Lisboa	0.71	356430.81	17.59	133
Cachoeira do Bom Jesus	0.70	380013.49	23.99	388
Estreito	0.69	360298.16	29.05	269
Vargem Pequena	0.67	416408.20	42.42	18

Bairro	**R^2**	**RMSE**	**MAPE**	**Observações**
Daniela	0.66	257050.53	12.36	15
Agronômica	0.65	436443.38	18.11	649
Jardim Atlântico	0.64	340714.88	26.95	253
Canasvieiras	0.64	397837.46	23.80	356
Armação do Pântano do Sul	0.62	385897.32	25.40	27
Saco Grande	0.61	305242.40	23.21	104
Coloninha	0.61	325530.45	30.91	76
Ponta das Canas	0.61	610054.76	28.52	28
Canto	0.59	289004.65	21.96	407
Itaguaçu	0.58	427727.93	25.31	54
Cacupé	0.57	430971.00	17.15	192
Ratones	0.52	565876.54	29.34	9
Bom Abrigo	0.51	472492.78	31.03	73
Ingleses do Rio Vermelho	0.48	399598.62	38.63	1775
Lagoa da Conceição	0.47	599580.21	24.71	179
Campeche	0.47	460092.15	21.81	1226
Praia Brava	0.47	466893.02	19.50	53
Abraão	0.43	277438.33	23.22	157
Rio Tavares	0.39	424470.13	20.91	267
Ribeirão da Ilha	0.37	333061.80	28.50	245
Jurerê	0.37	595242.23	22.12	323
Morro das Pedras	0.36	430840.63	22.33	80
Sambaqui	0.31	666064.25	28.09	35
Santa Mônica	0.28	429885.74	15.63	61
Jurerê Internacional	0.23	687991.08	24.94	236
Vargem Grande	0.23	487470.58	49.02	35
Pântano do Sul	0.22	468081.49	29.06	61
Pantanal	0.15	392400.18	28.65	73
Carianos	0.03	442978.03	31.47	73
Costeira do Pirajubaé	-0.01	344297.93	37.29	12
São João do Rio Vermelho	-0.06	470382.20	65.35	327
Capoeiras	-0.24	362831.33	36.31	396
Saco dos Limões	-0.67	357321.11	32.64	103
Monte Verde	-0.87	320743.50	26.89	54
Monte Cristo	-1.73	253590.58	28.86	19
Vargem do Bom Jesus	-2.74	407997.33	64.99	26
Tapera da Base	-8.23	565863.78	78.27	21

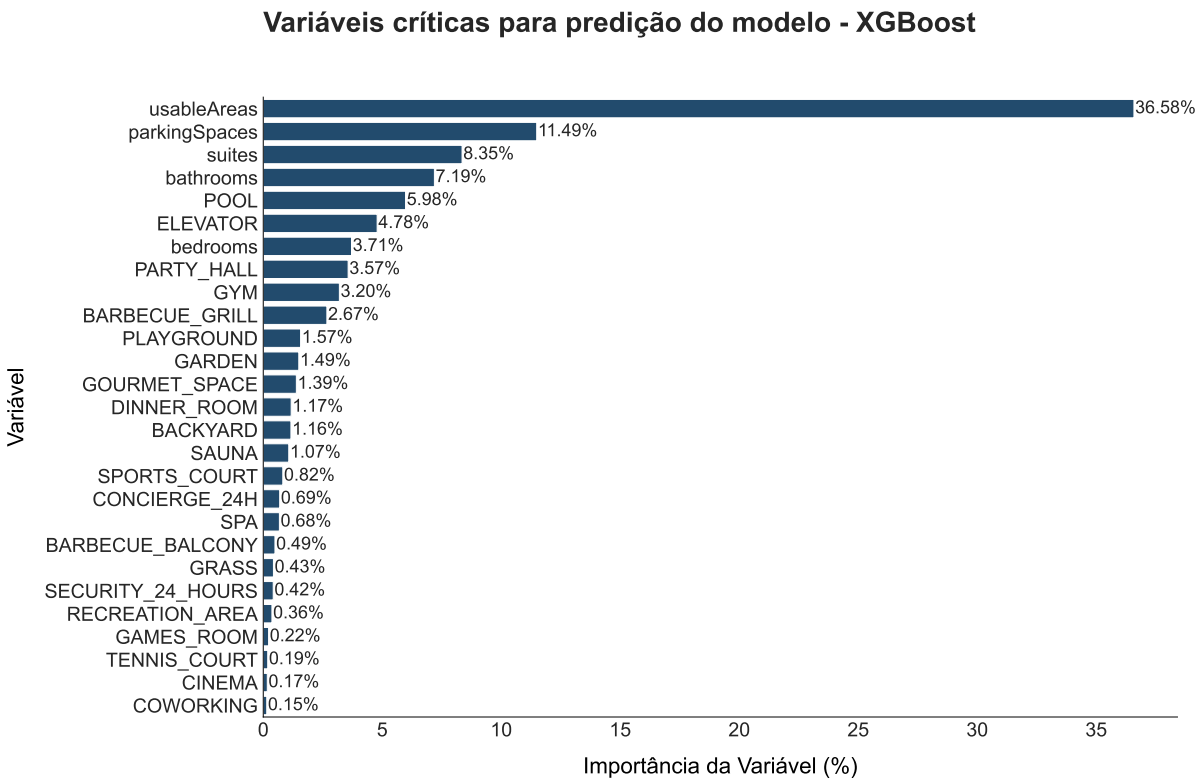
Bairro	**R^2**	**RMSE**	**MAPE**	**Observações**
José Mendes	-8.81	842188.32	119.36	4

Analisando os resultados específicos para cada bairro, é evidente que o desempenho do modelo varia significativamente entre eles. Bairros como Itacorubi (R^2 de 0,81), Córrego Grande (R^2 de 0,80), Coqueiros (R^2 de 0,79) e Trindade (R^2 de 0,78) destacam-se com valores de R^2 relativamente altos, indicando que o modelo genérico consegue capturar bem as variáveis que influenciam os preços dos imóveis nessas regiões. Estes bairros também apresentam valores de RMSE e MAPE razoáveis, corroborando a eficácia do modelo nessas áreas. Por outro lado, há bairros onde o desempenho do modelo é significativamente inferior, como Capoeiras (R^2 de -0,24), Saco dos Limões (R^2 de -0,67) e Monte Cristo (R^2 de -1,73). Nessas regiões, o modelo genérico não consegue capturar adequadamente as variáveis que influenciam os preços dos imóveis, resultando em valores negativos de R^2, o que indica que o modelo está performando pior do que uma média simples dos preços.

Os resultados globais do modelo genérico, com um R^2 de 0,70, sugerem que, enquanto o modelo é relativamente eficaz em uma visão geral, ele enfrenta desafios significativos em capturar as peculiaridades de bairros com características muito específicas ou menos representadas no conjunto de dados. A variação substancial de desempenho entre os bairros indica que a inclusão de variáveis adicionais ou a segmentação do modelo para regiões específicas poderia melhorar a precisão das previsões.

Para compreender em maior profundidade o comportamento do modelo com melhor perfomance (XGBoost ajustado), recorreu-se à biblioteca SHAP (*SHapley Additive exPlanations*). Através da análise SHAP, foi possível identificar as variáveis críticas para a previsão, ou seja, aquelas que mais influenciaram na determinação do valor previsto para o preço dos imóveis (Figura 27).

Figura 27 – Principais variáveis para predição do modelo - XGBoost



Fonte: Autor (2024)

Observa-se que a variável de maior influência na previsão do preço é a área útil, o que está alinhado com o comportamento observado no mercado imobiliário, onde o preço por metro quadrado (m²) e a localização são os dois atributos mais relevantes na precificação de um imóvel. Além disso, as demais variáveis, como garagem, suítes, banheiros e quartos, embora apresentem uma contribuição significativa em termos de percentagem, estão correlacionadas de forma indireta com a área útil. Isso ocorre devido ao fato de que um imóvel com mais quartos, por exemplo, naturalmente demanda uma área útil maior. Essa inter-relação entre as variáveis reflete a complexidade e a multifacetada natureza dos fatores que influenciam os preços dos imóveis.

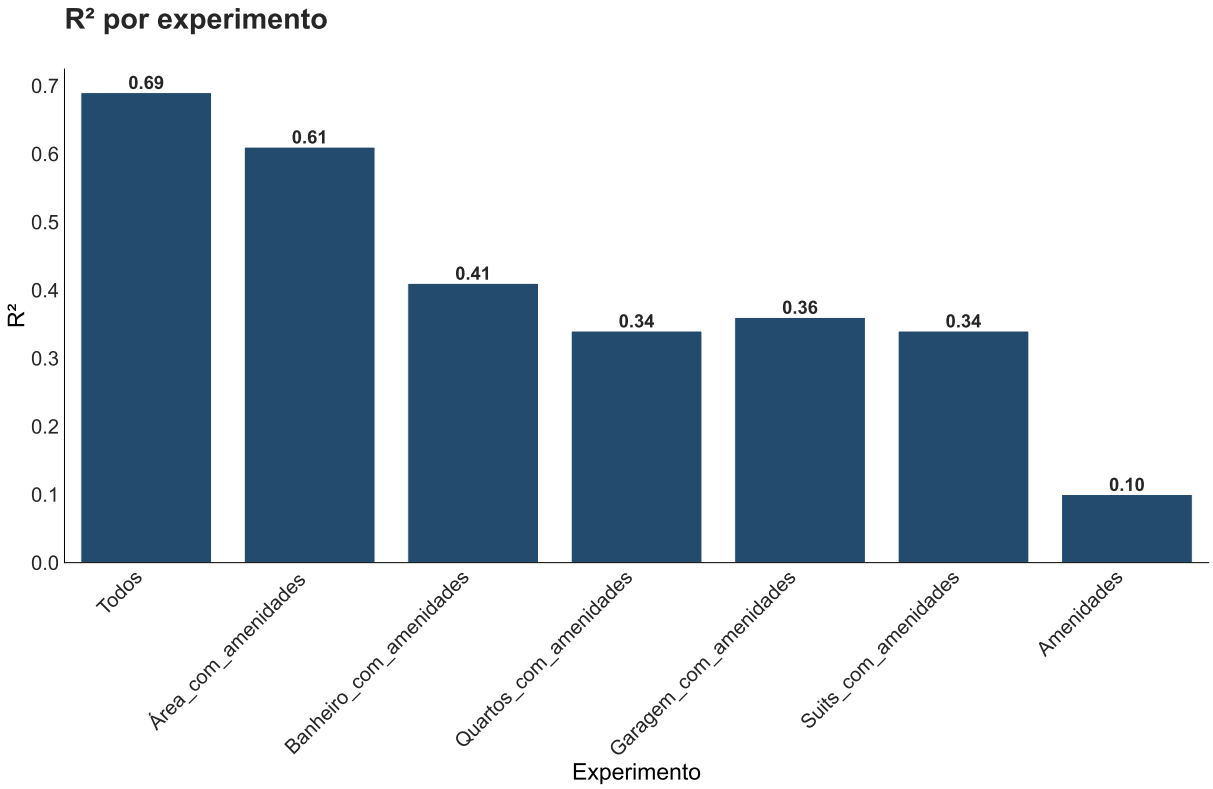
Afim de avaliar melhor como cada atributo afeta a perfomance do modelo, retreinou-se o modelo XGBoost otimizado para diferentes combinações de *features*, os quais podem ser resumidos a seguir:

- **Experimento 1**: Treino/Teste usando apenas a coluna área útil e amenidades.
- **Experimento 2**: Treino/Teste usando apenas a coluna banheiro e amenidades.
- **Experimento 3**: Treino/Teste usando apenas a coluna quartos e amenidades.
- **Experimento 4**: Treino/Teste usando apenas a coluna garagem e amenidades.

- **Experimento 5**: Treino/Teste usando apenas a coluna suites e amenidades.
- **Experimento 6**: Treino/Teste usando somente as amenidades.

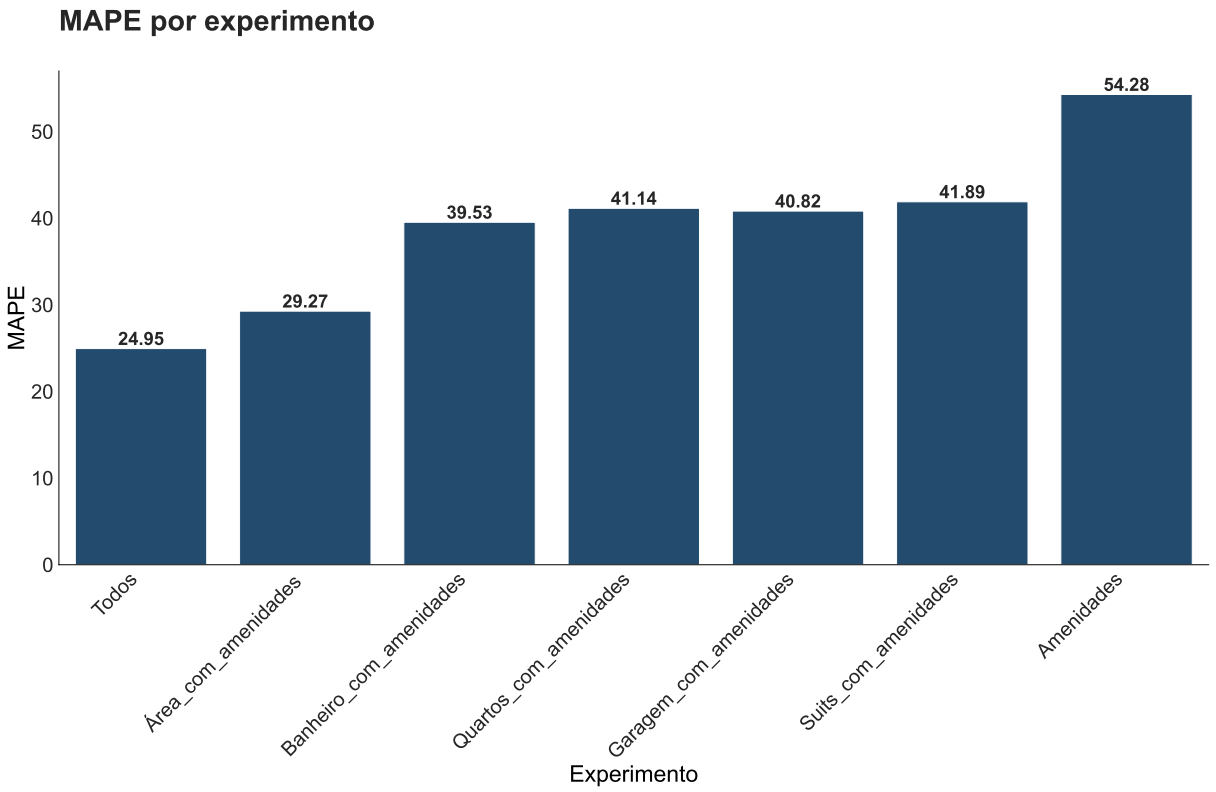
Os resultados de cada experimento podem ser vistos nas figuras 28 e 29, na qual se comparou a perfomance de R^2 de cada experimento com o melhor modelo de XGBoost otimizado (Coluna "Testes").

Figura 28 – R^2 para diferentes combinações de *features*



Fonte: Autor (2024)

Figura 29 – MAPE para diferentes combinações de *features*



Fonte: Autor (2024)

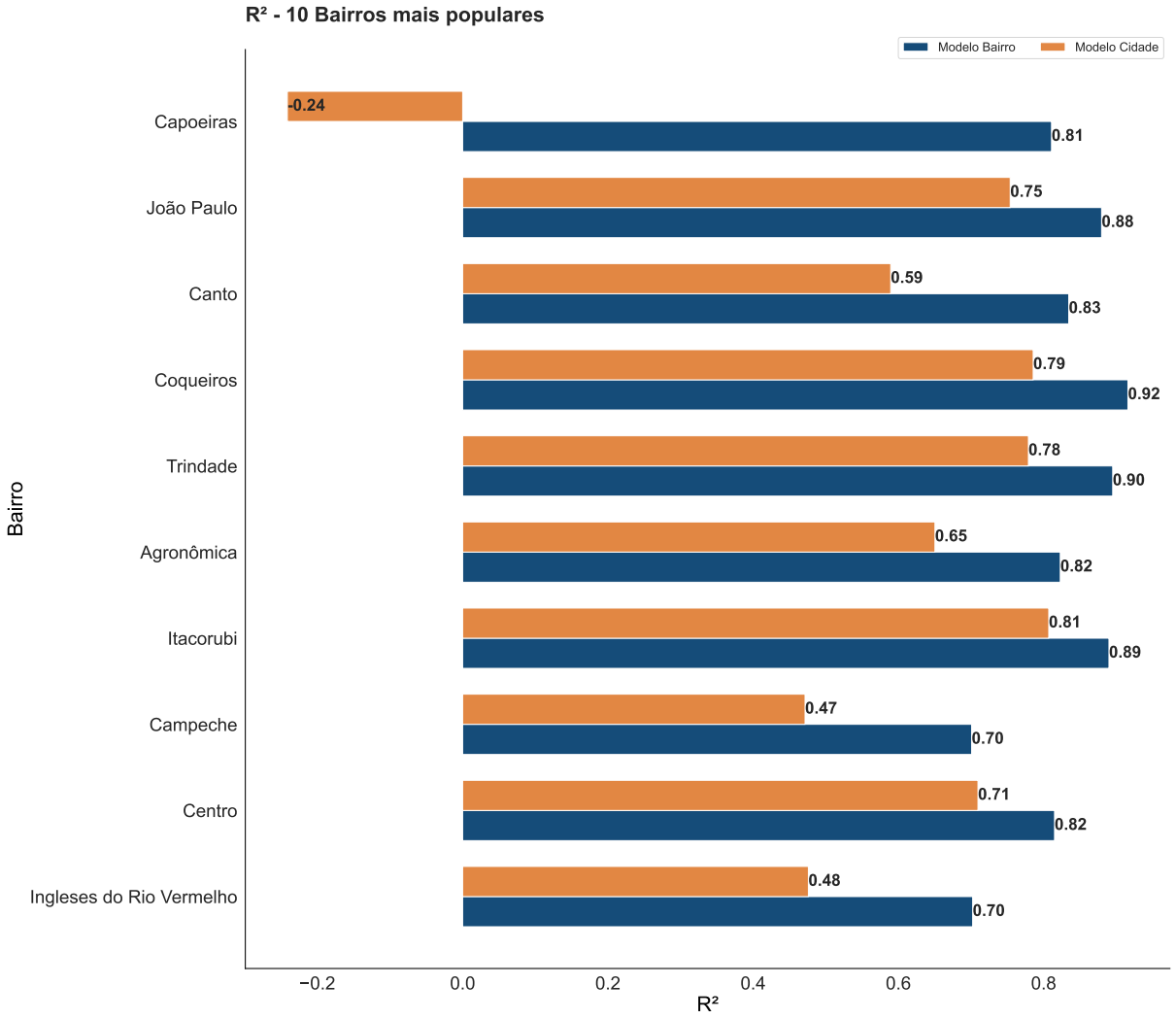
Os resultados indicam que variáveis específicas, como a área com amenidades e os banheiros com amenidades, possuem uma capacidade preditiva significativamente maior para a precificação de imóveis. No entanto, a combinação de todas as amenidades não necessariamente melhora o modelo, possivelmente devido à redundância das informações e à multicolinearidade. Isso não implica que as amenidades sejam irrelevantes, mas elas devem ser consideradas como fatores secundários que podem ajustar o valor final, em vez de componentes principais do modelo.

Com base nos resultados anteriores, surgiram dois questionamentos em relação ao comportamento do modelo. O primeiro questionamento diz respeito à performance instável do modelo em diferentes regiões. Uma hipótese para essa instabilidade é a existência de diferenças significativas entre os bairros, as quais não foram adequadamente capturadas pelos dados de treinamento. O segundo questionamento está relacionado ao volume de dados e seu impacto na performance do modelo.

Para investigar essas hipóteses, realizou-se uma análise comparativa adicional. Inicialmente, treinou-se o modelo XGBoost otimizado utilizando apenas os dados de bairros especificos. Para isso, utilizou-se os dez bairros com maior quantidade de observações, sendo eles: Ingleses do Rio Vermelho, Centro, Campeche, Itacorubi, Agronômica, Trindade, Coqueiros, Canto, João Paulo e Capoeiras. A comparação do desempenho entre os 10 modelos regionais com o modelo genérico pode ser visto na

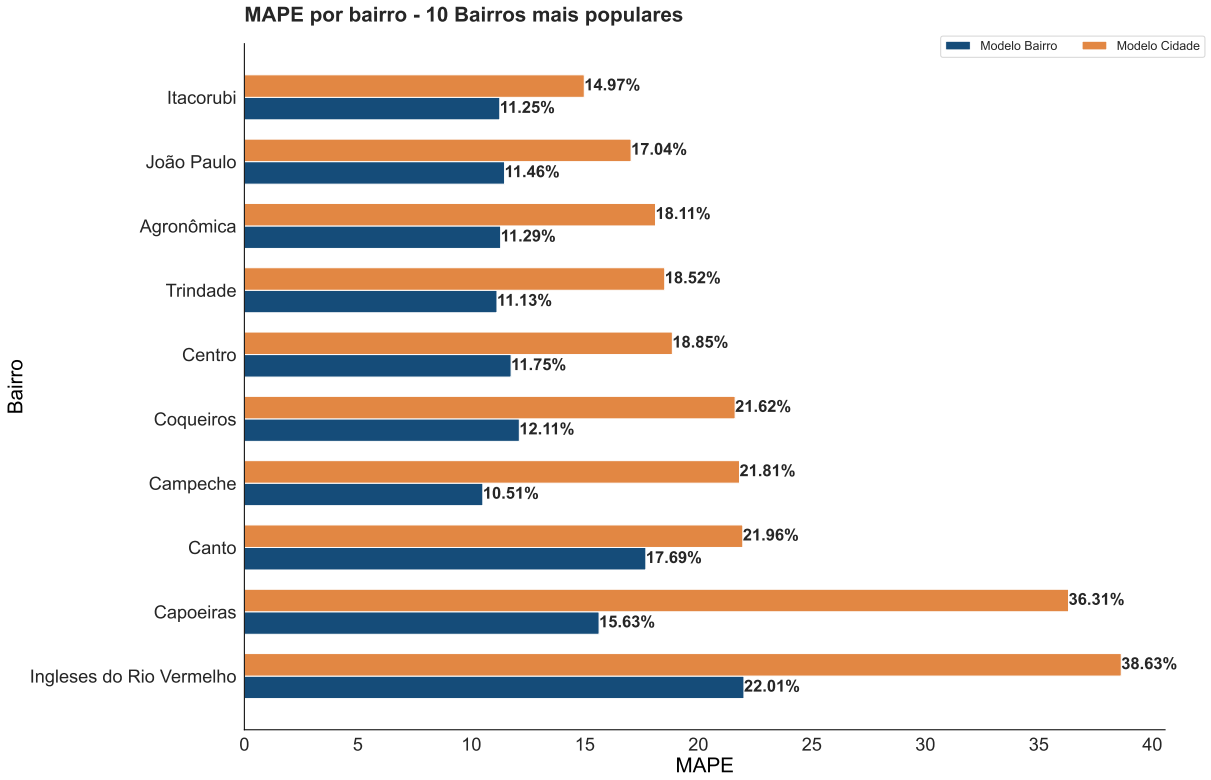
figura 30.

Figura 30 – Modelo regional vs modelo cidade - R^2.



Fonte: Autor (2024)

Figura 31 – Modelo regional vs modelo cidade - MAPE

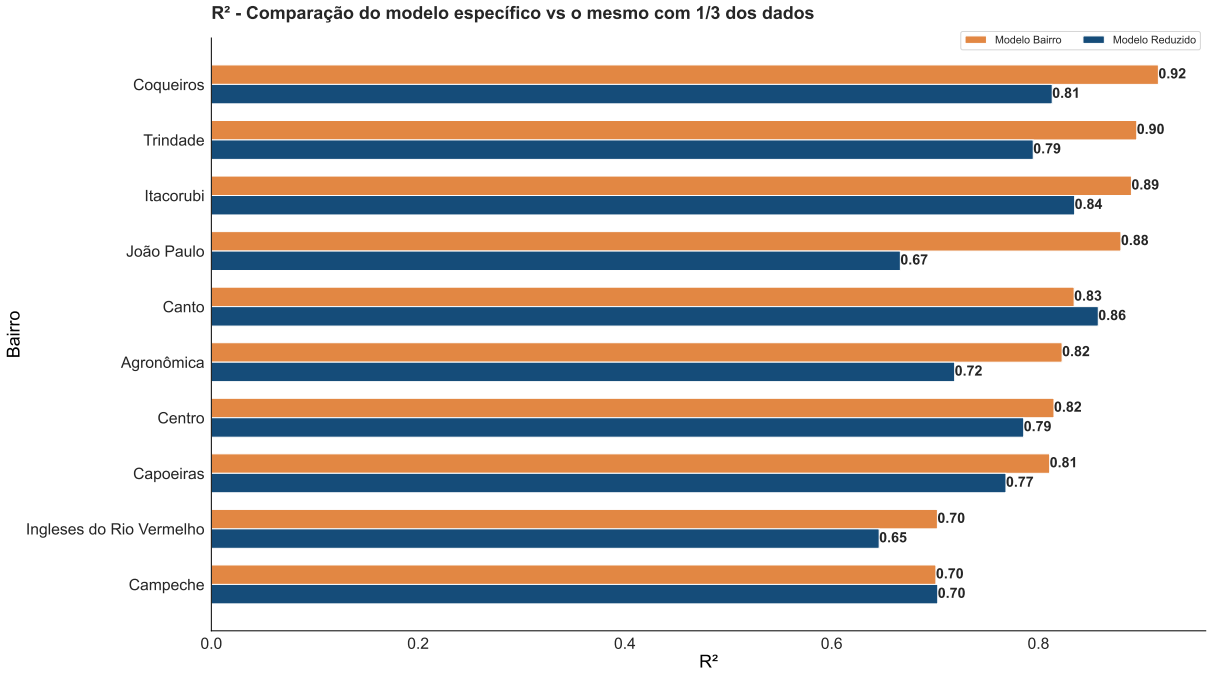


Fonte: Autor (2024)

A análise dos resultados revelou que, na maioria dos casos, o modelo de regressão treinado com dados específicos de bairros apresentou um desempenho superior ao modelo treinado com dados de toda a cidade. Isso destaca a importância das características locais na precificação de imóveis. Esses resultados sugerem que as características específicas de cada bairro são melhor capturadas quando o modelo é treinado exclusivamente com os dados daquela região, levando a uma melhoria substancial na precisão das previsões de preços de imóveis. Isso corrobora a hipótese de que a instabilidade na performance do modelo pode ser atribuída a diferenças significativas entre os bairros, que não são adequadamente representadas quando se utiliza um modelo treinado com dados de toda a cidade.

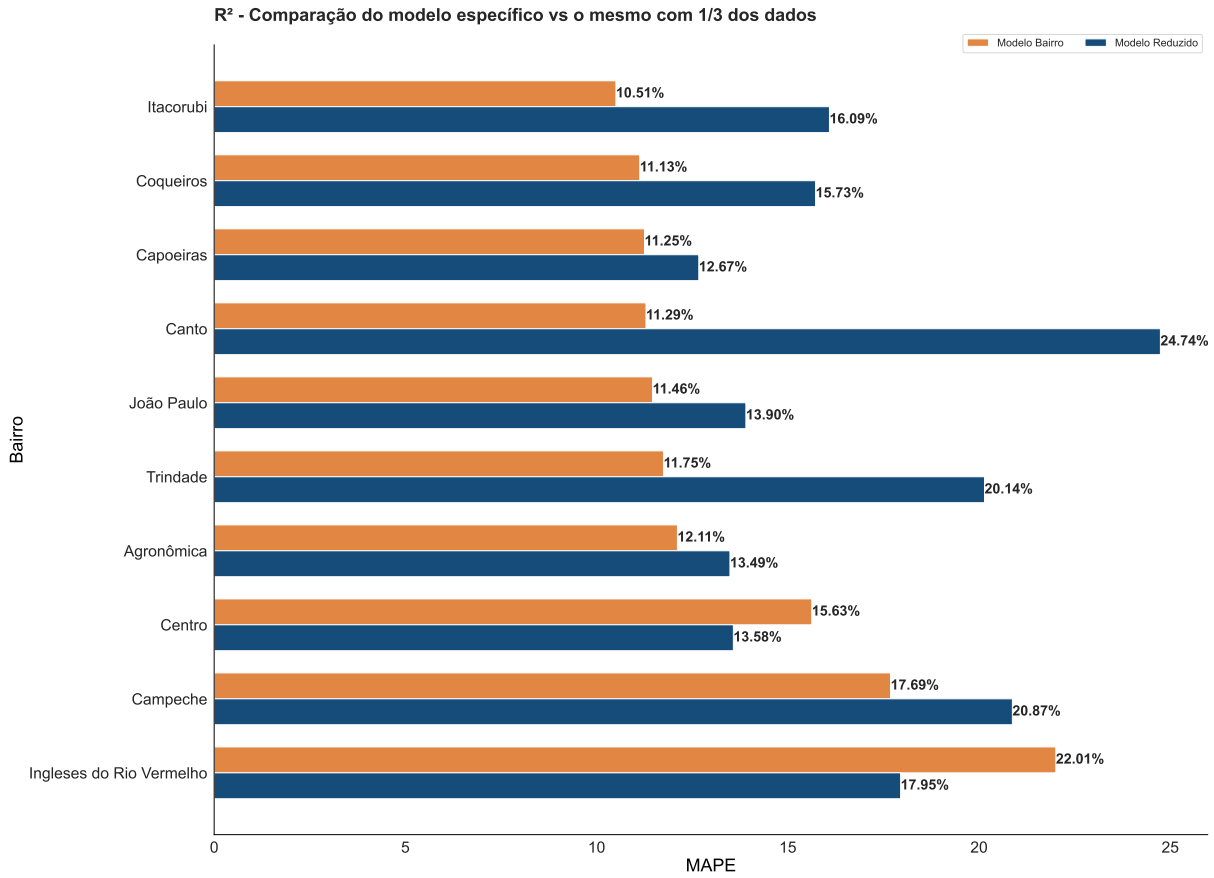
Por fim, retreinou-se o modelo genérico utilizando apenas um terço dos dados disponíveis, afim de se validar a segunda hipótese a respeito da importância do volume de dados para o modelo de precificação. O gráfico apresentado na figura 32 compara o desempenho do modelo específico para cada bairro com o desempenho do mesmo modelo quando treinado com apenas um terço dos dados disponíveis. Cada modelo foi treinado dez vezes selecionando 1/3 da amostra aleatoriamente, e disso se utilizou a média para análise. Esta comparação foi realizada para verificar a importância do volume de dados.

Figura 32 – Influência do volume de dados na perfomance do modelo - R^2



Fonte: Autor (2024)

Figura 33 – Influência do volume de dados na perfomance do modelo - MAPE



Fonte: Autor (2024)

Observa-se que, em todos os bairros analisados, o modelo específico treinado com a totalidade dos dados apresenta um desempenho superior ao modelo treinado com dados reduzidos. Por exemplo, em Coqueiros, o R^2 do modelo específico é 0,92, enquanto o do modelo com dados reduzidos é 0,81. Similarmente, em Trindade, o R^2 cai de 0,90 no modelo completo para 0,79 no modelo com um terço dos dados.

Esses resultados sugerem a importância do volume de dados para a performance do modelo. Um maior volume de dados permite que o modelo capture melhor as variáveis e as nuances específicas de cada bairro, resultando em previsões mais precisas. A redução no volume de dados diminui a capacidade do modelo de aprender e generalizar corretamente, levando a um desempenho inferior, conforme evidenciado pela redução dos valores de R^2 e aumento do MAPE. A análise reforça a hipótese de que a disponibilidade de dados ricos e abrangentes é fundamental para a precisão das previsões em estudos de precificação imobiliária.

4.6 CASOS DE USO

Para verificar a performance do modelo de precificação em novos dados, foram realizados testes que demonstraram como o modelo pode ajudar a estimar o preço dos imóveis. Esses testes envolveram a aplicação do modelo em amostras do conjunto de dados que não foram utilizadas durante o treinamento, permitindo avaliar sua capacidade de generalização. O objetivo foi observar se o modelo conseguiu prever os preços dos imóveis com um grau de precisão aceitável, fornecendo estimativas que se aproximam dos valores reais de mercado.

A figura 34 mostra alguns casos de uso do melhor modelo (XGboost otimizado) para amostras aleatórios dos dados de treino. Nelas, inseriu-se as variáveis usadas para o treinamento como área útil, número de quartos, banheiros, garagens, suites e lista de amenidades. A coluna bairro foi adicionada para representação, porém a mesma não foi usada como uma *feature* no treinamento, para evitar um possível sobreajuste do modelo.

Figura 34 – Caso de uso 1

	Neighborhood	Usable Areas	Suites	Bathrooms	Parking Spaces	Real Price	Preço Estimado pelo Modelo	Difference (%)	Amenities
1	Trindade	48	1.0	1.0	1.0	530000	568207.25	7.21%	GYM, BARBECUE_GRILL, PARTY_HALL, GOURMET_SPACE, DINNER_ROOM, RECREATION_AREA, ELEVAT
2	Jardim Atlântico	141	3.0	4.0	2.0	1402000	1568989.0	11.91%	BARBECUE_GRILL, POOL, PLAYGROUND, GOURMET_SPACE, SPORTS_COURT, SAUNA, CINEMA, ELEVAT
3	Trindade	68	1.0	2.0	1.0	838916	776362.94	-7.46%	GYM, BARBECUE_GRILL, PARTY_HALL, PLAYGROUND, GOURMET_SPACE, ELEVATOR

Fonte: Autor (2024)

Figura 35 – Caso de uso 2

	Neighborhood	Usable Areas	Suites	Bathrooms	Parking Spaces	Real Price	Preço Estimado pelo Modelo	Difference (%)	Amenities
1	Cachoeira do Bom Jesus	140	1.0	2.0	2.0	1450001	1401411.9	-3.35%	BARBECUE_GRILL, ELEVATOR
2	João Paulo	231	3.0	4.0	4.0	1900000	1985030.0	4.48%	BARBECUE_GRILL, PARTY_HALL, PLAYGROUND, ELEVATOR
3	Itacorubi	270	3.0	5.0	2.0	2300000	2300932.5	0.04%	BARBECUE_GRILL

Fonte: Autor (2024)

Figura 36 – Caso de uso 3

	Neighborhood	Usable Areas	Suites	Bathrooms	Parking Spaces	Real Price	Preço Estimado pelo Modelo	Difference (%)	Amenities
1	Itacorubi	76	1.0	2.0	1.0	927000	864042.75	-6.79%	BACKYARD
2	Centro	296	2.0	5.0	3.0	2130000	2467410.5	15.84%	PARTY_HALL, PLAYGROUND, GARDEN, GOURMET_SPACE, ELEVATOR
3	Ingleses do Rio Vermelho	75	1.0	2.0	1.0	855000	807656.8	-5.54%	

Fonte: Autor (2024)

Nas figuras 34, 35 e 36 podemos ver a precificação estimada pelo modelo, na coluna "Preço estimado pelo Modelo". Também calculou-se a diferença entre o preço real e o preço estimado, em percentagem (coluna *"Difference"*). Além de avaliar a precisão das previsões, os testes também ajudaram a identificar possíveis áreas de melhoria no modelo.

5 CONCLUSÃO

O objetivo deste estudo foi propor um modelo de precificação de imóveis para a região de Florianópolis, com base nos dados coletados por meio de *web scraping* da plataforma Viva Real para Janeiro de 2024. Para conduzir a pesquisa, adotaram-se as etapas do procedimento de ciência de dados. Essas etapas incluem a coleta de dados, limpeza e preparação dos dados, análise exploratória, treinamento e otimização dos modelos de *machine learning* e avaliação dos modelos. Ao longo do processo, foram utilizados três modelos de *machine learning* para a tarefa de previsão dos preços.

A fase inicial envolveu o desenvolvimento de um algoritmo de *web scraping* para extrair dados da plataforma imobiliária Viva Real. Esse algoritmo foi projetado para acessar as páginas da *web*, identificar as informações relevantes sobre os imóveis disponíveis em Florianópolis e extrair esses dados de forma automatizada para uma base de dados. Por meio desse processo de coleta de dados, obteve-se 78.960 observações contendo 32 atríbutos.

Após coletar os dados, avançamos para a etapa de pré-processamento, cujo objetivo foi manipular e selecionar as variáveis mais representativas para auxiliar no aprendizado do modelo. Concluída esta fase, obtivemos um conjunto de dados com 45.365 observações e 29 atributos, representando 51 diferentes bairros de Florianópolis.

Em seguida, foram treinados três diferentes modelos preditivos: *Lasso Regression*, *Random Forest* e *XGBoost*, com as devidas otimizações de hiperparâmetros, e os resultados foram avaliados utilizando validação cruzada para três métricas: R^2, RMSE e MAPE. O XGBoost alcançou o melhor resultado com um R^2 de 0.70 e MAPE de 24.12%. Observou-se que a otimização de hiperparâmetros não trouxe ganhos muito significativos para os modelos, tendo um acréscimo de apenas 3% para o XGBoost e 7% para o *Random Forest*. Analisando os resultados deste melhor modelo nos dados de teste de todos os bairros, observou-se uma performance errática entre os bairros. Isso sugere a possibilidade de existir alguma variável não incluída nos dados que tenha um peso maior do que os atributos usados para treinamento.

Além disso, utilizou-se do método SHAPLEY para identificar as características importantes do melhor modelo. Esse procedimento foi fundamental para compreender o impacto de cada atributo na previsão dos preços dos imóveis e possibilitar a interpretação do modelo. Neste caso, observou-se que as variáveis mais importantes na predição foram, *usabableAreas* (área útil) com 36%, parkingSpaces (vagas de garagem) 11.8%, suites - 8.75%, bathrooms (banheiro) - 8.38% e POOL (presença de piscina) - 5.98%. Tais resultados reforçam o comportamento esperado do mercado, cujo m^2 é o atributo de maior relevância na precificação de um imóvel.

Como sugestão para futuras pesquisas, recomenda-se a expansão da coleta de

dados para o treinamento dos modelos. Embora se reconheça as limitações relacionadas à disponibilidade de dados imobiliários, espera-se que esses obstáculos diminuam gradualmente à medida que os dados imobiliários se tornem mais acessíveis digitalmente. Outra sugestão é considerar o uso de outras variáveis que possam melhor refletir as preferências do consumidor, incluindo tanto variáveis geográficas como proximidade a locais de interesse, praias, shoppings e escolas, bem como características específicas do imóvel as quais não estavam disponíveis nos anúncios, como idade do prédio, tempo disponível no mercado e estado de conservação.

REFERÊNCIAS

ABRAMO, Pedro. A dinâmica imobiliária: elementos para o entendimento da espacialidade urbana. **Cadernos IPPUR/UFRJ**, n. 3, p. 47–70, 1988.

ALMAQBALI, Iqtibas Salim Hilal *et al.* Web Scrapping: Data Extraction from Websites. **Journal of Student Research**, 2019.

ARRAES, Ronaldo A.; SOUSA FILHO, Edmar de. Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso. **Economia aplicada**, v. 12, p. 289–319, 2008.

BALARINE, Oscar Fernando Osorio. **Determinação do impacto de fatores sócio-econômicos na formação do estoque habitacional em Porto Alegre**. [*S.l.*]: Edipucrs, 1996.

BALDOMINOS, Alejandro *et al.* Identifying real estate opportunities using machine learning. **Applied sciences**, MDPI, v. 8, n. 11, p. 2321, 2018.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, p. 5–32, 2001.

CAN, Ayşe. GIS and spatial analysis of housing and mortgage markets. **Journal of Housing Research**, v. 9, n. 1, p. 61–86, 1998.

CHAI, Tianfeng; DRAXLER, Roland R. Root mean square error (RMSE) or mean absolute error (MAE). **Geoscientific Model Development Discussions**, v. 7, n. 1, p. 1525–1534, 2014.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. *In:* PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [*S.l.*: *s.n.*], 2016. P. 785–794.

CHOY, Lennon HT; HO, Winky KO. The use of machine learning in real estate research. **Land**, MDPI, v. 12, n. 4, p. 740, 2023.

CRAVEN, B. D.; ISLAM, Sardar MN. Ordinary least-squares regression. **The SAGE Dictionary of Quantitative Management Research**, p. 224–228, 2011.

D'AMATO, Valéria *et al.* Pension schemes versus real estate. **Annals of Operations Research**, v. 299, p. 797–809, 2021.

DIN, Allan; HOESLI, Martin; BENDER, Andre. Environmental variables and real estate prices. **Urban Studies**, v. 38, n. 11, p. 1989–2000, 2001.

EVANS, Alan W. The property market: ninety per cent efficient? **Urban Studies**, v. 32, n. 1, p. 5–29, 1995.

FIKER, José. **Manual de avaliações e perícias em imóveis urbanos**. [*S.l.*]: Oficina de Textos, 2001.

FIPEZAP+ (SÃO PAULO). RELATÓRIO ÍNDICE RESIDENCIAL INFORME MARÇO 2024. [*S.l.*: *s.n.*], 2024. `https://downloads.fipe.org.br/indices/fipezap/fipezap-202403-residencial-venda.pdf`. Acesso em: 20 maio 2024.

GARSON, G. David. **Statnotes: Topics in multivariate analysis**. Acesso em 02 de fevereiro de 2021. 2009. Disponível em: `https://faculty.chass.ncsu.edu/garson/PA765/statnote.htm`.

GÉRON, Aurélien. **Hands-on machine learning with scikit-learn and tensorflow: Concepts, Tools, and Techniques to build intelligent systems**. [*S.l.*: *s.n.*], 2017.

GOLDSTEIN, Benjamin A; NAVAR, Ann Marie; CARTER, Rickey E. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. **European heart journal**, Oxford University Press, v. 38, n. 23, p. 1805–1814, 2017.

GONZÁLEZ, Marco Aurélio Stumpf. **Aplicação de técnicas de descobrimento de conhecimento em bases de dados e de inteligência artificial em avaliação de imóveis**. 2002. Tese (Doutorado).

GONZÁLEZ, Marco Aurélio Stumpf; FORMOSO, Carlos Torres. Análise conceitual das dificuldades na determinação de modelos de formação de preços através de análise de regressão. **Engenharia Civil–UM (Universidade do Minho)**, v. 8, p. 65–75, 2000.

GUAN, Jian *et al.* Analyzing massive data sets: an adaptive fuzzy neural approach for prediction, with a real estate illustration. **Journal of Organizational Computing and Electronic Commerce**, v. 24, n. 1, p. 94–112, 2014.

HACHCHAM, Aymane. **XGBoost: Everything You Need to Know**. [*S.l.*: *s.n.*], 2024. Acesso em: 7 jul. 2024. Disponível em: `https://neptune.ai/blog/xgboost-everything-you-need-to-know`.

HASTIE, Trevor; TIBSHIRANI, Robert; WAINWRIGHT, Martin. Statistical learning with sparsity. **Monographs on statistics and applied probability**, v. 143, n. 143, p. 8, 2015.

HILLEN, Judith. Web scraping for food price research. **British Food Journal**, v. 121, n. 12, p. 3350–3361, 2019.

HO, Winky KO; TANG, Bo-Sin; WONG, Siu Wai. Predicting property prices with machine learning algorithms. **Journal of Property Research**, Taylor & Francis, v. 38, n. 1, p. 48–70, 2021.

IBM. **What Is Random Forest?** [*S.l.*: *s.n.*], 2024. Acesso em: 7 jul. 2024. Disponível em: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems..

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina: uma abordagem estatística**. [*S.l.*]: Rafael Izbicki, 2020.

KHDER, Moaiad Ahmad. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. **International Journal of Advances in Soft Computing & Its Applications**, v. 13, n. 3, 2021.

KOBAYASHI, Mei; TAKEDA, Koichi. Information retrieval on the web. **ACM Computing Surveys (CSUR)**, v. 32, n. 2, p. 144–173, 2000.

KOK, Nils; KOPONEN, Eija-Leena; MARTÍNEZ-BARBOSA, Carmen Adriana. Big data in real estate? From manual appraisal to automated valuation. **The Journal of Portfolio Management**, v. 43, n. 6, p. 202–211, 2017.

KUHN, Max; JOHNSON, Kjell *et al.* **Applied predictive modeling**. [*S.l.*]: Springer, 2013. v. 26.

LI, Sheng *et al.* Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (XGBoost-HPM). **Land**, v. 10, n. 5, p. 533, 2021.

LINNE, Mark R; KANE, M Steven; DELL, George. **A guide to appraisal valuation modeling**. [*S.l.*]: Appraisal Institute, 2000.

LUNDBERG, Scott M; LEE, Su-In. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MAGRI ZAGHI, Lucca *et al.* Modelo de previsão de preços de imóveis na cidade de Florianópolis/SC a partir de técnicas de machine learning. Florianópolis, SC., 2023.

MATOS, Débora; BARTKIW, Paula Izabela Nogueira. **Introdução ao mercado imobiliário**. Curitiba: IFPRE-tec, 2013.

MORO, Matheus Fernando. **Modelo híbrido de séries temporais para previsão de demanda do mercado imobiliário de São Paulo**. 2017. Tese (Doutorado).

MULLAINATHAN, Sendhil; SPIESS, Jann. Machine learning: an applied econometric approach. **Journal of Economic Perspectives**, v. 31, n. 2, p. 87–106, 2017.

NARULA, Subhash C; WELLINGTON, John F; LEWIS, Stephen A. Valuating residential real estate using parametric programming. **European journal of operational research**, v. 217, n. 1, p. 120–128, 2012.

PARK, Byeonghwa; BAE, Jae Kwon. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. **Expert Systems with Applications**, v. 42, n. 6, p. 2928–2934, 2015.

PEDERSEN, Anne Marie B.; WEISSENSTEINER, Alex; POULSEN, Rolf. Financial planning for young households. **Annals of Operations Research**, v. 205, p. 55–76, 2013.

PERSSON, Emil. Evaluating tools and techniques for web scraping, 2019.

PROBST, Philipp; WRIGHT, Marvin N.; BOULESTEIX, Anne-Laure. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: data mining and knowledge discovery**, v. 9, n. 3, e1301, 2019.

PROSKURIN, Oleksandr. **Bagging in Financial Machine Learning: Sequential Bootstrapping. Python example**. [*S.l.*: *s.n.*], 2024. Acesso em: 7 jul. 2024. Disponível em: `https://hudsonthames.org/bagging-in-financial-machine-learning-sequential-bootstrapping-python/`.

RAMPINI, Luca; RE CECCONI, Fulvio. Artificial intelligence algorithms to predict Italian real estate market prices. **Journal of Property Investment & Finance**, v. 40, n. 6, 2022.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **Information**, v. 11, n. 4, p. 193, 2020.

ROBINSON, Ray. **Housing economics and public policy**. [*S.l.*]: Springer, 1979.

ROZEMBERCZKI, Benedek *et al.* The shapley value in machine learning. **arXiv preprint arXiv:2202.05594**, 2022.

SHAPLEY, Lloyd S *et al.* A value for n-person games. Princeton University Press Princeton, 1953.

SHARMA, Hemlata; HARSORA, Hitesh; OGUNLEYE, Bayode. An Optimal House Price Prediction Algorithm: XGBoost. **Analytics**, MDPI, v. 3, n. 1, p. 30–45, 2024.

SHEN, Helen. Interactive notebooks: Sharing the code. **Nature**, v. 515, n. 7525, p. 152–152, 2014.

SKANSI, Sandro. **Introduction to Deep Learning: from logical calculus to artificial intelligence**. [*S.l.*]: Springer, 2018.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 58, n. 1, p. 267–288, 1996.

VALIER, Agostino. Who performs better? AVMs vs hedonic models. **Journal of Property Investment & Finance**, v. 38, n. 3, p. 213–225, 2020.

WOLPERT, David H. The lack of a priori distinctions between learning algorithms. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . ., v. 8, n. 7, p. 1341–1390, 1996.

WU, Jia *et al.* Hyperparameter optimization for machine learning models based on Bayesian optimization. **Journal of Electronic Science and Technology**, Elsevier, v. 17, n. 1, p. 26–40, 2019.

XU, Kevin; NGUYEN, Hieu. Predicting housing prices and analyzing real estate market in the Chicago suburbs using Machine Learning. **arXiv preprint arXiv:2210.06261**, 2022.

ANEXO A – REPOSITÓRIO PARA O CÓDIGO

Projeto Imobiliário