Universidade de Brasília
Departamento de Estatística

Técnicas de aprendizado de máquinas aplicadas a metodologia de Engenharia de Avaliação: Projeção do valor locativo de um imóvel

Victor do Nascimento Rezende

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Victor do Nascimento Rezende

Técnicas de aprendizado de máquinas aplicadas a metodologia de Engenharia de Avaliação: Projeção do valor locativo de um imóvel

Orientador: Leandro Tavares Correia

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Resumo

A Engenharia de Avaliação é uma área de grande importância para o estudo e obtenção do valor locativo de um imóvel. Neste trabalho, com o intuito de trazer maior rigor estatístico as etapas deste processo e também para a escolha das variáveis a serem utilizadas na criação do modelo, foi utilizado a metodologia já usual da Engenharia de Avaliação com a aplicação de técnicas de aprendizado de máquinas. O Banco de dados foi recebido e reorganizado, desde as variáveis referente a especificação e localização dos dados até as variáveis quantitativas e qualitativas a serem utilizadas nas análises resultando em um banco com 189 dados e 25 variáveis. Posteriormente, foi realizado o processo de seleção de variáveis com o auxílio de técnicas de aprendizado de maquinas, por meio da técnica K-Fold. Nesta análise utilizou-se também as técnicas de Árvore de Regressão e Florestas Aleatórias para se ampliar as análises referente ao conjunto de variáveis. Por fim, a junção de técnicas possibilitou a melhor escolha das variáveis, através das análises paramétricas e não paramétricas e a melhor modelagem e definição de um modelo final, através do SISDEA, possibilitando a criação de um modelo com graus de precisão e fundamentação 3, dentro da Norma, sendo estes os melhores padrões e consequentemente o melhor modelo para a conclusão do valor locativo de um imóvel que se enquadre na comparação com os dados comparativos da amostra.

Palavras-chave: Engenharia de Avaliação, método comparativo, regressão, $R\$/AP$, valor locativo, modelo, variáveis, SISDEA, K-Fold, aprendizado de máquinas, projeção.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Sabe-se da importância da regressão linear múltipla como uma técnica para análise simultânea de efeitos que várias variáveis independentes causam em uma variável dependente. Essa ferramenta permite analisar não só o valor obtido final para a variável dependente, mas também a influência que as demais variáveis tem na formação do valor final.

Com isso, a regressão múltipla torna-se um grande aliado quando se pensando em análise de mercado e com um grande potencial para análise imobiliária ainda mais sabendo que no Brasil são escassos os estudos relacionados a esse tema em que muitas as vezes a determinação do valor, locativo ou venal, de um imóvel, sejam eles comerciais ou residenciais é feita através de inferência estatística ou por uma simples comparação de valor de mercado (DANTAS, 2005).

Embora exista essa deficiência, o Brasil mostrou-se como um dos países mais desenvolvidos nos últimos anos na área da avaliação de imóveis. Esse desenvolvimento tornou-se possível por conta da introdução a metodologia cientifica como ferramenta essencial para o trabalho avaliatório e também da aparição de diferentes sistemas de tratamento de dados (DANTAS, 2005).

Temos então que a avaliação, especificamente imobiliária, é a análise de um ou mais fatores econômicos, escolhidos de acordo com as características do estudo especifico e com uma data determinada, tendo como suporte a análise de dados relevantes (ABUNAHMAN, 2008). Assim, há no Brasil uma área especifica de trabalho, relativamente nova, chamada Engenharia de Avaliação que por meio de um processo de análise de dados e estatísticas, com a utilização de sistemas específicos, tem por objetivo definir o valor teórico de um imóvel, seja locativo ou venal.

Para esse tipo de trabalho há a norma padrão regulamentada sendo ela a NBR 14653, que tem por objetivo garantir que o modelo de regressão, que será utilizado, seja bom o suficiente para que se torne possível a realização da projeção do valor do bem a ser avaliado. Os resultados obtidos desse trabalho são de grande importância, tendo em vista que a maioria das transações de imóveis são feitas através de financiamento e também ao grande quantitativo de contratos locativos que são firmados e a conclusão obtida por esse trabalho é utilizada como uma garantia tanto para quem irá comprar ou alugar o imóvel, quanto para a instituição que irá ceder o crédito ou receber o valor de uma locação.

Para trazer um maior rigor e novas interpretações estatísticas para do processo de Engenharia de Avaliação, que por sua vez só pode ter o relatório final assinado por Engenheiros e Arquitetos, será aplicado conceitos e métricas nas principais etapas deste trabalho afim de se ter um critério e um aprofundamento estatístico ainda maior.

Este trabalho apresenta uma revisão geral no banco de dados, sendo essa a primeira etapa do processo, bem como o acréscimo de variáveis com o intuito de melhorar e enriquecer as análises realizadas. Posteriormente, foi realizado uma análise exploratória para possibilitar a observação da distribuição dos dados para as devidas variáveis utilizadas. Por fim, foram utilizados métricas e parâmetros estatísticos, afim de se obter a melhor combinação de variáveis para possibilitar a criação e o ajuste do melhor modelo possível com o auxílio do Software SISDEA, software popularmente utilizado na Engenharia de Avaliação.

Além do método paramétrico de regressão linear, já incorporado neste processo, foi aplicado técnicas de aprendizado de máquinas como a validação cruzada K-Fold e métodos não paramétricos como Árvores de Regressão e de florestas aleatórias, afim de se construir um melhor modelo para que seja obtido melhores conclusões para os valores de um imóvel.

2 Revisão de Literatura

2.1 Mercado Imobiliário

Temos que o mercado é um local no qual agentes econômicos transacionam bens uns com os outros onde se é trocado bens de serviço por uma quantia monetária ou até mesmo por outros bens. Com isso, tem-se então que o mercado é o conjunto destes agentes, interagindo entre si, com a finalidade de comprar e vender os seus produtos ou serviços. Dentro de uma situação de mercado, temos o equilíbrio sendo ditado pela lei da oferta e procura, sendo a oferta a quantidade oferecida de um determinado produto ou bem e procura sendo a demanda por estes (MATOS; BARTKIW, 2013).

O mercado, na forma como vemos atualmente, sofre influencias de variáveis macroeconômicas, como é citado por (KREMER, 2008), tendo como principais fatores influentes a renda, produtos nacionais, emprego e desemprego, entre outros. Também é possível ver essa influência no quesito microeconômico como a oferta e demanda, a ampla concorrência e também os impostos, pontos estes abordados também por (MATOS; BARTKIW, 2013).

> A estrutura do mercado depende, fundamentalmente, de três variáveis: número de empresas que compõem esse mercado; tipo de produto gerado nesse mercado; e a existência ou inexistência de barreiras de entrada para novas empresas nesse mercado (KREMER, 2008).

Sabendo dessa influência, temos também que um dos pontos principais do mercado é o ambiente em que ele está inserido e ao qual ele atende. Com ele podemos definir o tipo de cliente, as suas necessidades e as suas prioridades. O aumento do mercado tem crescimento proporcional com a quantidade de pessoas interessadas por ele, assim como o aumento também das suas exigências. No fim, o produto ou serviço que melhor atendes a necessidade dessa procura, receberam melhor destaque e consequentemente terão uma maior procura (MATOS; BARTKIW, 2013).

No Brasil há um mercado com uma grande relevância, o Mercado Imobiliário. Esse mercado é responsável pelas negociações e transações de bens imóveis podendo ser residências como apartamentos, casas e também comerciais como prédios, lojas, salas, terrenos entre outros. Este tipo de mercado contempla também toda a fase administrativa de um imóvel como compra, venda, aluguel, financiamento. Como este é um mercado que lida com a negociação de patrimônios com valores altos, tem-se então a necessidade de uma especialização para aqueles que trabalham na área, levando assim um desenvolvimento da área cada dia maior.

Temo como definição do mercado imobiliário, abordado por (SOUSA, 2006):

> O Mercado imobiliário é o mecanismo social de coordenação das decisões individuais e localização e uso do solo urbano, de forma que desta coordenação surge uma cidade cujo solo urbano é (ou deveria ser) utilizado de forma mais eficiente. Cabe a esse mercado conciliar a liberação de ação individual com a eficiência dos recursos da sociedade. Temos, assim, a metáfora do mercado imobiliário como sendo a 'mão invisível urbana' que promove a cidade eficiente.

Uma das consequências, positivas, que este mercado trás é o desenvolvimento do espaço urbano afetando assim diretamente na qualidade de vida da sociedade. Este mercado é responsável por gerar, diretamente e indiretamente, um grande número de empregos pelos seus serviços, como, por exemplo, incorporação imobiliária, corretagem, publicidade e sistema financeiro habitacional levando assim a um crescimento da economia local (KREMER, 2008).

Embora seja complicado o acesso a informação sobre os parâmetros do mercado imobiliário no Brasil, (KREMER, 2008), trouxe alguns comparativos entre os anos de 2001 a 2006 onde temos que o número de faturamento cresceu de R$34 bilhões em 2001 para R$59,6 bilhões de reais em 2006, para o parâmetro de investimentos, temos que a evolução foi de R$ 21,7 bilhões para R$45,9 bilhões de reais, respectivamente entre 2001 a 2006. Embora esses dados sejam de anos anteriores, da para se ter uma ideia do crescimento do Brasil nesta área.

Em Brasília, assim como no Brasil, este mercado se desenvolve cada vez mais, chegando a alguns meses bater os melhores parâmetros do País. Para fins de comparação, temos a pesquisa do índice de Velocidade e Venda (IVV), criado pelo Sindicato da Indústria da Construção Civil do Distrito Federal (Sinduscon-DF) em parceria com a Associação de Empresas do Mercado Imobiliário (Ademi-DF). Esta pesquisa tem como objetivo gerar índices que permitam o acompanhamento da comercialização dos imóveis novos, garantido uma análise conjuntural do mercado local. Com isso, podemos então comparar que a média de venda mensal de imóveis aumentou de 240 em 2016 para 427 em 2021, comparando o mês de dezembro de ambos os anos. Vemos também um aumento no valor ponderado do m^2 no Distrito Federal, aumentando de R$7.920,00 em dezembro de 2016, para R$ 10.676,00 no mesmo período do ano de 2021.

Por outro lado, se observamos os parâmetros desta pesquisa apenas para os imóveis comercias, vemos que o ano de 2021 não apresenta os melhores resultados, comparando com anos anteriores até 2015. Isso nos deixa claro que, embora o Mercado Imobiliário seja totalmente presenta em nossa sociedade, é um tipo de mercado muito volátil e que apresenta diversas alterações, sendo necessário então, como já abordado, cada dia mais estudos sobre este tipo de mercado por meio dos agentes que trabalham nele e tornando assim cada vez mais ao desenvolvimento desta área.

2.2 Engenharia de Avaliação

Temos em nossa sociedade uma grande concentração de bens imóveis e que fazem parte não só da vida pessoal de cada um, mas também estão presentes em órgãos públicos e privados, tendo assim então, uma grande relevância. Com essa importância e também sabendo do alto volume destes bens, surge então a necessidade de qualificarmos o valor de mercado deste bem, dando assim suporte a futuras decisões relativas a disposição e uso deste imóvel (ABUNAHMAN, 2008).

A avaliação pode ser definida como um processo e resultado de se obter valores definidos de um imóvel, assim como sua utilidade e também a possibilidade de venda (ABUNAHMAN, 2008). O Engenheiro Rubens Alves Dantas completa definindo e especificando que é uma área da engenharia responsável por reunir as áreas da engenharia e arquitetura com as ciências socias, exatas e da natureza, com o intuito de poder determinar, baseado em técnicas, o valor de um bem de qualquer natureza (DANTAS, 2005). Sergio Abunahman traz ainda que, a confiabilidade do valor definido é parte responsável da competência e integridade básicas do avaliador, da disponibilidade de dados possíveis de se obter e também da qualidade e habilidade com que esses dados obtidos foram tratados e analisados (ABUNAHMAN, 2008).

> Avaliação é, pois, uma aferição de um ou mais fatores econômicos especificamente definidos em relação a propriedades com data determinada tendo como suporte a análise de dados relevantes (ABUNAHMAN, 2008).

No Brasil, Luiz Carlos Berrini como sendo um dos principais e primeiros escritores brasileiros sobre a Engenharia de Avaliação, tendo lançado em 1941 o seu primeiro livro sobre o assunto porém, muito antes, entre 1918 e 1919, tem-se relatos de publicações feitas em revistas no estado de São Paulo (DANTAS, 2005).

A avaliação tem por objetivo refletir a tendencia do mercado derivado da análise dos dados coletados para este trabalho. É necessário também, que não haja nenhuma interferência pessoal do avaliador pois, a avaliação deve ser uma opinião sustentável. Isso se deve, pois, a principal finalidade deste trabalho é concluir o valor do bem para que seja possível utiliza-lo em futuras decisões (ABUNAHMAN, 2008).

Juntamente com o crescimento e desenvolvimento do mercado imobiliário, a área de Engenharia de Avaliações acompanhou também esse processo, sendo possível notar cada vez mais a aplicação de metodologias no processo deste trabalho. Atualmente, acredita-se que o Brasil seja um dos países com maior relevância nesta área no mundo (DANTAS, 2005).

A necessidade e também a utilização dos resultados obtidos por este tipo de trabalho, são listados por Sergio Abunahman (ABUNAHMAN, 2008), tais como:

1) Possibilidade de uma transação de propriedade como: Auxiliar compradores na decisão de um preço de oferta, auxiliar vendedores na decisão de um preço de venda justo, estabelecer bases de permuta de propriedades e auxiliar nas decisões em caso de fusões e incorporações de Empresas.

2) Financiamento e Crédito como: Garantia de empréstimos sob forma de hipoteca;Fornecer informações que auxiliem investidores na decisão quanto a compra de bens imóveis hipotecados, ações ou outro tipo de apólice; Estabelecer parâmetros para a emissão ou endosso de empréstimos com base no bem possuído.

3) Justa Indenização nos Casos de Desapropriação como: Concluir sobre o valor da propriedade, antes da possível desapropriação; Concluir sobre o valor da propriedade pós o processo de desapropriação.

4) Tomada de Decisão Sobre Bens Imóveis como Analisar o mercado afim de se identificar e quantificar qual o mais provável para o bem imóvel assim como os seus prazos; Analisar as tendencias do mercado em relação ao uso pretendido de um terreno; Concluir sobre as metas propostas de investimentos para aquele bem imóvel.

5) Base Para Taxações (Impostos) como: A partir da distinção dos valores da benfeitoria com o valor do terreno, calcular os índices de desvalorização aplicáveis; Determinar impostos sobre heranças ou doações.

6) Aplicações Secundárias como: Auxiliar na relação entre seguradores e cliente a determinação do valor real do prêmio correspondente.

7) Justo Valor Locacional como: Possibilitar a conclusão do valor locatício justo tanto para o locatário quanto para o locador; Fornecer subsídios ao Juízo para aplicação de sentenças nas Ações Renovatórias e Revisionais.

Então, a avaliação imobiliária procura poder concluir sobre o valor de mercado do imóvel, seja ele venal ou locativo. (ABUNAHMAN, 2008) destaca três definições para o conceito de valor de mercado:

> Valor de Mercado é o maior preço em termos de dinheiro que o imóvel pode ter uam vez posto à venda, abertamente, por um tempo razoável para encontra comprador, o qual deverá ter conhecimento de todos os usos, propósitos e utilidades, para que ele, comprador tenha capacidade de utilizar o imóvel (ABUNAHMAN, 2008).

ou, ainda,

> Valor de Mercado é o preço pago por um comprador desejoso de compra, mas não forçado, a um vendedor desejoso de vender, nas também não compelido, tendo ambos plenos conhecimentos da utilidade da propriedade transacionada (ABUNAHMAN, 2008).

Por fim, Abunahman traz um conceito que é próximo ao emitido pelo Engenheiro Mexicano Enrique Lira Montes de Oca (ABUNAHMAN, 2008):

> Valor de Mercado é o preço que um vendedor está disposto a aceitar, e um comprador a pagar, ambos perfeitamente bem informados e dentro de circunstancias normais, objetivos e subjetivas, para um determinado bem (ABUNAHMAN, 2008).

E para se obter este valor, é necessário seguir a determinada metodologia, qual seja: Realizar uma pesquisa de vendas ou alugueis de propriedades comparáveis a que se deseja avaliar, Fazer um comparativo entre as propriedades utilizadas na pesquisa e o imóvel a ser avaliado e pesquisar a tendência central ou a média ponderada dos resultados obtidos para chegar, finalmente ao Valor (ABUNAHMAN, 2008).

Método Comparativo de Dados de Mercado

Para a realização deste trabalho há a norma (NBR-14.653-1, 2019) complementada pela (NBR-14.653-2, 2011), de acordo com a ABNT- Associação Brasileira de Normas Técnicas, que direciona as metodologias a serem aplicadas no trabalho. Nesta norma, temos a descrição dos métodos para a identificação do valor e todos os parâmetros que devem ser seguidos, assim como também a descrição para o enquadramento do trabalho em graus de precisão e fundamentação.

A (NBR-14.653-1, 2019) cita os tipos de métodos como os métodos comparativos direto de dados de mercado, método involutivo e o método evolutivo. Traz também, a descrição sobre os métodos de capitalização de renda, da quantificação de custo, método comparativo direto de custo e o método para indicadores de viabilidade da utilização econômica do empreendimento.

Este trabalho será baseado exclusivamente no método comparativo de dados de mercado. Segundo determinação das normas da ABNT, este é o método mais indicado e que apresenta os resultados mais exatos. (NETO, 2021) descreve esse método como:

> O Método comparativo direto de dados de mercado, conforme definição dada pela norma ABNT NBR 14.653-1, "Identifica o valor de mercado do bem por meio de tratamento técnico dos atributos dos elementos comparáveis, constituinte da amostra", devendo ser utilizado sempre que possível. Este método contempla as seguintes etapas: planejamento da pesquisa, identificação das variáveis do modelo, levantamento de dados de mercado e tratamento dos dados.

Neste método, busca-se criar uma amostra, já que na maioria dos casos não é possível obter-se a informação de todos os bens ofertados em um determinado mercado, cujos valores médios forneçam estimativas do valor médio entre todos os bens que compões a população. Com isso, tomando como base a amostra de dados que sejam diretamente comparáveis ao bem a ser avaliando, tem-se por objetivo conseguir chegar a uma conclusão do valor mais provável de mercado deste imóvel (NETO, 2021).

3 Metodologia

3.1 Método de Avaliação de Imóveis

Como já descrito acima, as avaliações seguem a norma (NBR-14.653-2, 2011) está fazendo parte da ABNT – Associação Brasileira de Normas Técnicas, em que há descrito todos os passos necessários para a realização deste trabalho. A norma cita os principais métodos utilizados para a avaliação de imóveis, sendo eles o Método Comparativo de Dados de Mercado, Método Evolutivo, Método Involutivo, Método da Capitalização da Renda.

Este trabalho, terá como foco o Método Comparativo de Dados de Mercado pelo fato de este ser o método mais utilizado e devido a citação da norma que diz que este deve ser utilizado sempre que possível.

O Método comparativo de dados de mercado é definido pela norma ABNT (NBR-14.653-1, 2019) como sendo "o valor de mercado do bem por meio de tratamento técnico dos atributos dos elementos comparáveis, constituinte da amostra". Para realização deste trabalho tem-se as seguintes etapas: realização da pesquisa, identificação das variáveis do modelo, levantamento de dados de mercado e o tratamento dos dados.

A utilização deste método consiste na utilização de um banco de dados de bens que se assemelham ao bem que se deseja avaliar, caracterizados pelas principais variáveis que possam levar a formação de valor do mesmo. Mas, sabendo que não é possível a utilização de todos os dados disponíveis em oferta no mercado é utilizada uma amostra, cujo valor médio fornece uma estimativa do valor médio entre todos os bens que compõe a população.

Na Engenharia de Avaliações, o preço de um imóvel pode ser definido como uma expressão monetária dos dados de mercado em oferta ou efetivamente transacionados, sendo representado por:

$$P = f(\beta_1 x_1, \beta_2 x_2, ..., \beta_n x_,) + \epsilon,$$

onde:

- f é o indicativo da forma funcional.
- P é o preço do bem.
- $x_1, x_2, ...x_n$ são as características ou atributos relacionados a questões estruturais (físicas), de localização e aspectos econômicos e temporais.

- $\beta_1, \beta_2, ..., \beta_n$ são os parâmetros a serem estimados.
- ϵ é erro nas estimativas realizadas.

3.2 Banco de Dados

Visando aplicar as técnicas deste trabalho em uma situação real de Engenharia de Avaliação, foi obtido um banco de dados utilizado por uma empresa do ramo que é composto por dados locativos de lojas em diversos empreendimentos como Shoppings, Mall's, Universidades e no Aeroporto, todos nas regiões administrativas de Brasília e também em duas cidades consideradas como entorno do Distrito Federal, Valparaiso e Luziânia. Essas informações são de grande importância pois tratam-se de dados de difícil acesso que raramente podem ser obtidos por meio de fontes comuns de busca de imóveis da internet e que foram obtidos ao longo de anos por meio de contatos pessoais feitos com gerentes prediais e corretores que trabalham especificamente com este ramo.

O banco de dados foi recebido com 189 dados e uma composição de 19 variáveis sendo as 5 primeiras variáveis padrões e que devem sempre constar no levantamento de dados, de acordo com a Norma (NBR-14.653-1, 2019). As demais variáveis foram escolhidas por serem fatores importantes e que com o conhecimento de mercado, foram consideradas como prováveis influentes na formação do valor locativo das lojas. Abaixo descrição das variáveis como foi passada pela empresa.

Variáveis

Abaixo relação das variáveis padrões, de acordo com a Norma (NBR-14.653-1, 2019) e (NBR-14.653-2, 2011) .

- Endereço: endereço ou principal identificação do imóvel;
- Complemento: complementos que possam ser relevantes as características do imóvel;
- Bairro: bairro em que o imóvel se localiza;
- Informante: nome do informante ou empresa que forneceu o dado;
- Telefone: telefone do informante ou empresa que forneceu o dado.

Variáveis independentes:

- AP: variável quantitativa contínua, representativa da área privativas em m^2 do imóvel;
- Ano: variável temporal, representativa do ano de coleta do dado;

- Fluxo: variável quantitativa discreta, representativa da quantidade de pessoas que frequentam o empreendimento por dia;
- Lojas: variável quantitativa discreta, representativa da quantidade de lojas presentes no empreendimento;
- Inaug: variável temporal, representativa do ano de inauguração do empreendimento;
- Público: variável qualitativa, representativa da renda do público que frequenta o imóvel;
- Plan=2: variável dicotômica, indicando se o empreendimento está localizado no Plano Piloto;
- Of=2: variável dicotômica, indicando a natureza mercadológica do dado;
- Mall/Posto=1: variável dicotômica, indicando se o imóvel está localizado em um Mall/Posto;
- R$/mês: variável quantitativa, indicando o valor em reais por mês do imóvel;
- Segmento: variável qualitativa, indicando o segmento referente ao imóvel;
- C Comer=2: variável dicotômica, indicativa do imóvel está ou não localizado em um Centro Comercial;
- Shop=2: variável dicotômica, indicativa do imóvel está ou não localizado em um Shopping;
- R$/mês/AP: variável quantitativa, indicando o valor em reais por mês por área privativa.

Ajuste dos Dados

Em primeira vista aos dados, percebeu-se uma falta de padronização quanto a descrição das primeiras variáveis tidas como padrões para a Norma (NBR-14.653-1, 2019) sendo estas, as principais formas de demonstração da localização, tipologia e forma de obtenção destes dados.

Primeiramente a variável endereço foi padronizada de acordo com endereço cartorial utilizado, informação obtida através do site do geoportal disponibilizado pelo Governo do Distrito Federal, para que todas as lojas pertencentes a um mesmo empreendimento tivessem o endereçamento igualmente descritos. Para os dados das cidades de Luziânia e Valparaiso, foi utilizado a padronização usualmente utilizada pelo próprio empreendimento.

A variável complemento trazia em seu preenchimento o nome do empreendimento e em muitas ocasiões o nome ou tipologia da Loja, com isso foi então criado duas novas variáveis sendo elas empreendimento, com a descrição padronizada do nome do empreendimento ao qual a loja pertence e a variável loja, com a informação do nome ou tipologia da mesma.

A descrição da variável bairro em sua maioria trazia a informação da micro região ao qual o dado se encontrava como por exemplo SCS, sigla para Setor Comercial Sul, descrição dos bairros no caso das lojas localizadas em Valparaiso e em algumas ocasiões até a informação quanto ao pavimento que estava localizado a loja dentro do empreendimento então, essa variável foi substituída pela variável Região Administrativa e preenchida com a região ao qual o empreendimento encontra-se dentro do seu perímetro.

As demais variáveis foram reorganizadas para uma melhor visualização do banco de dados. Houve também a retirada da variável que fazia referencia a centro comercial, por entender que o critério do preenchimento da mesma não se fazia claro e incluída seis novas variáveis referente ao estacionamento privativo, renda per capita da região, grupo de renda per capita, população da região, variável indicativa de universidade e a tipologia do empreendimento, imaginando que alguma destas possa se mostrar influente na composição do valor locativo de uma loja.

Com isso, o novo banco de dados e o que será utilizado para todas as análises e conclusões, conta com os mesmos 189 dados porém, com 25 variáveis sendo as 6 primeiras compostas por informações que contemplem a exigência da Norma e as demais 19 variáveis com informações que de alguma forma possam trazer influencia na criação de um modelo capaz de predizer o valor loja comparativa aos dados que compõe o banco. Abaixo relação das variáveis do banco de dados final.

Variáveis Ajustadas

- Endereço: descrição completa do endereço de acordo com o registro cartorial obtido no site do geoportal do Governo do Distrito Federal e para as regiões fora do Distrito Federal, utilizado o endereço usualmente adotado pelo empreendimento;
- Empreendimento: descrição padronizada do nome do empreendimento como usualmente é utilizado pelo mesmo;
- Loja: nomenclatura ou tipologia da loja;
- Região Administrativa: região em que o empreendimento se encontra em seu perímetro.
- Informante: nome ou local ao qual a informação foi obtida;
- Telefone: telefone para contato com quem se obteve a informação;

- AP: área privativa da loja em metros quadrado;
- Estac. Privado=2: indicativo do empreendimento ter (=2) ou não (=1) a opção de estacionamento privado;
- Fluxo: referente ao fluxo médio diário de público no empreendimento.
- Lojas: quantidade total de lojas no empreendimento;
- Inaug: data de inauguração do empreendimento;
- Público: referente ao público alvo do empreendimento. Público de renda alta (=4), público de renda média alta (=3), público de renda média baixa (=2), público de renda baixa (=1);
- Renda Per Capita: renda per capita da região administrativa de acordo com PDAD de 2018. Para as regiões foram do Distrito Federal, foi utilizada a informação do censo de 2010 do IBGE;
- Grupo Renda: agrupamento em grupos de acordo com a classificação da PDAD de 2018.
- População: população estimava das regiões de acordo com PDAD de 2018. Para as regiões foram do Distrito Federal, foi utilizado a informação constante do censo de 2010 do IBGE;
- Plan=2: indicativa do imóvel está localizado no Plano Piloto (=2) ou não (=1).
- Segmento: referente ao segmento da loja no empreendimento;
- Shopp=2: indicativa do imóvel está localizado em Shopping (=2) ou nos demais empreendimentos (=1); Mall/Posto=2: indicativa do imóvel está localizado em Mall ou Posto (=2) ou nos demais empreendimentos (=1);
- Universidade=2: com indicação análoga as variáveis acima só que para Universidade (=2);
- Tipo: tipologia da loja em que temos, Aeroportos (=4), Shopping (=3), Universidades (=2) e Mall ou Postos de Combustível (=1);
- Ano: ano de coleta do dados;
- Ofe=2: natureza mercadológica do dado e que indica se a loja estava alugada (=1) no momento da coleta da informação ou se estava em oferta (=2).

E as duas variáveis possíveis de utilização como dependentes, sendo:

- R$/mês: Valor mensal em reais;
- R$/mês/AP: Valor do metro quadrado mensal em reais.

3.3 Análise Descritiva

Será realizado uma análise descritiva de todas a variáveis utilizadas para a criação do banco de dados com a finalidade de entender-se melhor o funcionamento e a distribuição das mesmas. Essa primeira análise nos permitira primeiramente entender quais variáveis visualmente aparentam ser importantes para a composição do modelo bem como a distribuição dos dados para cada uma delas.

Análise Bivariada

Por se tratar de um estudo de um banco de dados voltado ao valor locativo de imóveis localizado em conjuntos comerciais e universidades, será utilizado a análise bivariada que nos permitirá observar a distribuição dos dados entre as variáveis a serem utilizadas como independentes em relação variável a ser utilizada como dependente, de acordo com o modelo a ser criado.

Para o estudo bidimensional entre duas variáveis quantitativas, serão utilizados gráficos de dispersão para a análise visual e também será calculado a medida de dispersão de correlação de Pearson que testara a correlação estatística entre as variáveis independentes com a variável dependente bem como os pares das variáveis independentes. Este coeficiente tem um intervalo de valores entre +1 e -1 em que valores próximos a 0 são indicativas de uma não associação, já valores positivos próximos a 1 são indicativos de uma relação positiva, ou seja, à medida que o valor da variável aumenta o par que está sendo comparado também tem o seu valor aumentado e valores próximos a -1 são indicativos de uma relação negativa, com relação entre os pares se dando de moto contrário a relação anterior explicada, (BUSSAB; MORETTIN, 2010).

Sejam x_i e y_i valores das variáveis X e Y. A fórmula da correlação de Pearson se dará de acordo com a fórmula 3.3.1:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum(x_i - \overline{x})^2(\sum(y_i - \overline{y})^2)}}. \tag{3.3.1}$$

Para a análise entre uma variável qualitativa e a variável dependente quantitativa, será optado pela utilização do BoxPlot que nos permitira analisar visualmente as características de localização, dispersão, assimetria, comprimento da cauda, medidas discrepantes e principalmente a distribuição dos dados, de acordo com cada código utilizado na composição da variável qualitativa. A visualização do BoxPlot pode ser bem resumida

de acordo com a Figura 1.

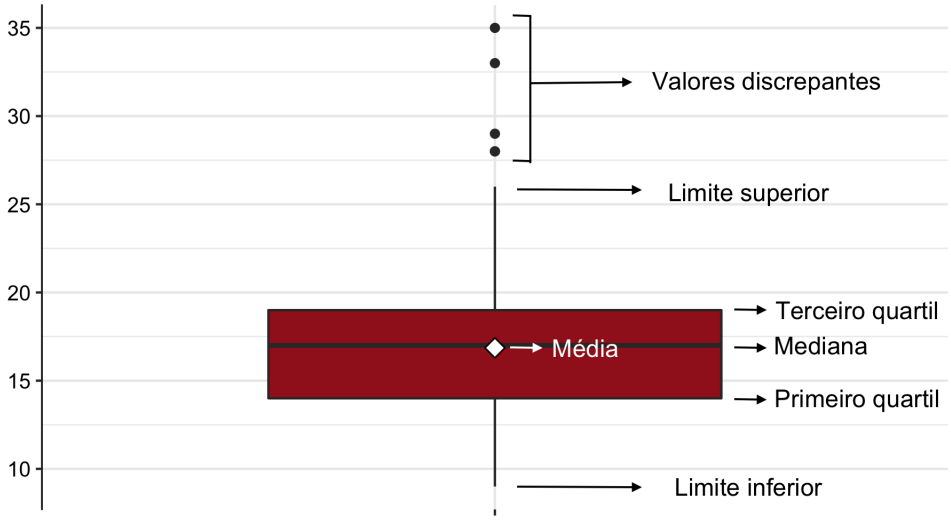

Figura 1: Análise do BoxPlot.

Fonte: ESTAT Consultoria.

A partir dessa etapa, utilizaremos o R, Software de tratamento estatístico e também o SISDEA um dos Softwares mais conhecidos para a realização do trabalho de Engenharia de Avaliação, criado pelo Engenheiro Civil e Mecânico, Antônio Pelli.

3.4 SISDEA

O SIDEA é um software para Avaliação de Imóveis Urbanos, Rurais e de Máquinas e Equipamentos, que possibilita a modelagem de dados através da Regressão Linear, Regressão Não Linear, RNA- Redes Neurais Artificias e DEA - Envelopamento de dados. Esse sistema nos permitirá o tratamento de dados e amostras do mercado imobiliário juntamente com a interpretação dos resultados estatísticos obtidos.

Para mais sobre o SISDEA, recomenda-se o acesso a pagina oficial da Pelli Sistemas Engenharias (NETO, 2022).

3.5 Regressão Linear Múltipla

O modelo de regressão linear múltiplo com p variáveis tem a forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i. \tag{3.5.1}$$

- Y_i é o valor da variável resposta para a i-ésima observação;

- $\beta_0, \beta_1, \ldots, \beta_p$ são parâmetros desconhecidos;
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas;
- ε_i são independentes e $\mathcal{N}(0, \sigma^2)$.

A função resposta para o modelo é

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{3.5.2}$$

Quando $p = 1$ o modelo é um modelo linear simples, da forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{3.5.3}$$

Outro caso especial é quando $p = 2$, o que traz uma superfície de resposta na forma.

E para estimar $\beta_0, \beta_1, \ldots, \beta_p$ é usado mínimos quadrados ordinários, em que as estimativas de Beta são:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.5.4}$$

Em que:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{3.5.5}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \tag{3.5.6}$$

3.6 Valor Predito

O valor predito é o valor que o modelo prediz, considerando os valores das variáveis explicativas, que é definido como:

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{3.6.1}$$

Intervalo de Predição

Considerando $\hat{Y}_n$ o valor predito de uma nova variável, a variável $\hat{Y}_n$ segue uma distribuição Normal com média $E(Y_n)$ e variância obtida pela formula 3.6.2:

$$\sigma^2\left[1+\frac{1}{n}+\frac{(X_0-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]. \tag{3.6.2}$$

Com isso, um intervalo de predição $(1-\alpha)$ para $E(Y_n)$ é dado pela equação 3.5.4:

$$IC(E(Y_n);1-\alpha)=\left(\hat{Y}_n-t_{(1-\frac{\alpha}{2};n-(p+1))}\sqrt{QMResA};\hat{Y}_n+t_{(1-\frac{\alpha}{2};n-(p+1))}\sqrt{QMResA}\right), \tag{3.6.3}$$

em que,

$$A=\left[1+\frac{1}{n}+\frac{(X_0-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right], \tag{3.6.4}$$

e também,

$$\sigma^2=QM_{Res}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}. \tag{3.6.5}$$

3.7 Modelos Nâo Paramétricos

Árvores de Regressão

Temos que uma Árvore de Decisão, voltada para a análise de regressão, consiste em uma metodologia não paramétrica, que pode nos possibilitar outros tipos de interpretações que não puderem ser vistas utilizando os métodos estritamente paramétricos.

A utilização da árvore para prever uma nova observação é feita do seguinte modo: começando pelo topo, verificamos se a condição descrita no topo (primeiro nó) é satisfeita. Caso seja, seguimos a esquerda. Caso contrário, seguimos a direita. Assim prosseguimos até atingir uma folha (IZBICKI; SANTOS, 2020). Como demonstrado na Figura 2:

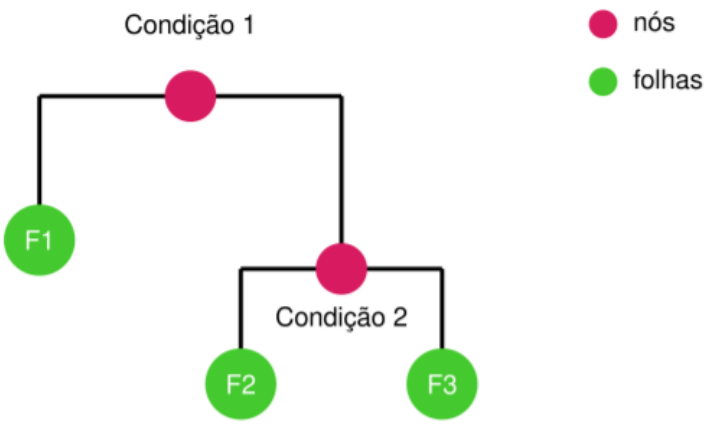

Figura 2: Exemplo de estrutura de uma árvore de regressão.

Fonte: (IZBICKI; SANTOS, 2020).

Formalmente, uma árvore cria uma partição do espaço das covariáveis em regiões distintas e disjuntas: $R_1, ..., Rj$ (IZBICKI; SANTOS, 2020). A predição para a resposta Y de uma observação com covariáveis x que estão em R_k é dada por:

$$g(x) = \frac{1}{|i : x_i \epsilon \mid R_k|_{i:x_i \epsilon R_k}} \sum y_i. \tag{3.7.1}$$

Então, observado a região em que a observação esta é calculada a média dos valores da variável resposta das amostras do conjunto de treinamento pertencentes àquela mesma região. Para a criação dessa arvore de criação, primeiro é construído uma árvore completa e consequentemente bastante complexa e depois é realizada a poda dessa arvore para que assim seja evitado o super ajuste.

É então utilizado o erro quadrático médio, para ser analisado cada árvore T, da seguinte forma:

$$P(T) = \sum_R \sum_{i:x_i \epsilon R} \frac{(y_i - \hat{y_R})^2}{n}, \tag{3.7.2}$$

em que $\hat{y_R}$ é o valor predito para a resposta de uma observação pertencente à região R. Encontrar T que minimize $P(T)$ é computacionalmente inviável (IZBICKI; SANTOS, 2020). Para que seja encontrado uma árvore com erro quadrático médio baixo que consistira na criação de divisões binárias recursivas, utiliza-se de uma heurística. Com isso, para essa escolha de partição, busca-se, dentro todas as covariáveis x_i e cortes t_1, aquela combinação que leva a partição (R_1, R_2) com predições de menor erro quadrático. Após, busca-se particionar R_1 ou R_2 em regiões menores e para seguir com o processo da escolha

da nova divisão, a mesma estratégia é utilizada. Então, este processo é repetido até que chegue a uma árvore com poucas observações em cada uma das folhas.

Florestas Aleatórias

Florestas aleatórias (BREIMAN, 2001) são métodos que contornam a limitação do baixo poder preditivo das árvores de decisão combinando diversas árvores para fazer uma predição para um mesmo problema.

Esse tipo de ajuste consiste em criar B árvores distintas e combinar seus resultados para melhorar o poder preditivo em relação a uma árvore individual. Para calcular a medida de importância de uma variável, primeiro se deve calcular o total da redução da soma de quadrados dos resíduos para todas as divisões que foram feitas com base nessa variável. Com isso, repetindo o processo para todas as árvores, será calculado então a média dessa redução que será utilizada como medida de importância para cada variável (IZBICKI; SANTOS, 2020).

3.8 Diagnóstico de Modelo

Temos na criação de modelo a obtenção do parâmetro de risco observado, que se observado é um estimador muito otimista. Uma maneira de solucionar este problema é dividir o conjunto de dados em duas partes, treinamento e validação (IZBICKI; SANTOS, 2020) e assim podemos então avaliar o erro quadrático médio do conjunto de validação .

Uma boa prática para escolher quais amostras serão utilizadas para compor o conjunto de treinamento e quais serão utilizadas para compor o conjunto de validação é fazê-lo aleatoriamente. Dessa forma, utiliza-se um gerador de números aleatórios para escolher quais amostras serão usadas para o treinamento e quais serão usadas para a validação. Esse procedimento evita problemas quando o banco de dados está previamente ordenado de acordo com alguma covariável (por exemplo, quem coletou o banco pode ter ordenado as observações em função de alguma variável) (IZBICKI; SANTOS, 2020). Esse procedimento de divisão dos dados para utilizar uma parte para a estimação do risco leva o nome de data splitting.

K-fold

O método de validação K-fold (IZBICKI; SANTOS, 2020) consiste na divisão dos dados aleatoriamente em k-folds(lotes) disjuntos com aproximadamente o mesmo tamanho. Sejam $L_1, ..., L_k \subset 1, ..., n$ os índices associados a cada um dos lotes. A ideia da validação pelo método de K-fold é criar k estimadores da função de regressão, $\hat{g}_{-1}, ..., \hat{g}_{-k}$, em que $\hat{g}_{-j}$ é criado usando todas as observações do banco menos aquelas do lote L_j.

4 Resultados

4.1 Análise Bivariada

Com o intuito de estudar as possíveis relações das variáveis em relação a variável dependente R$/mês/AP, inicialmente serão feitas análises descritivas bivariadas. É importante lembrar que essa primeira análise pode ser mascarada por diversas questões que de alguma maneira possam não ter sido contempladas.

Variáveis Quantitaivas

Inicialmente começaremos observado os gráficos de dispersão entre as variáveis quantitativas e a variável dependente R$/mês/AP, afim de poder estudar visualmente o comportamento das mesmas. Será analisado também, um gráfico de calor demonstrando a associação através do cálculo da correlação entre todos os pares dessas variáveis.

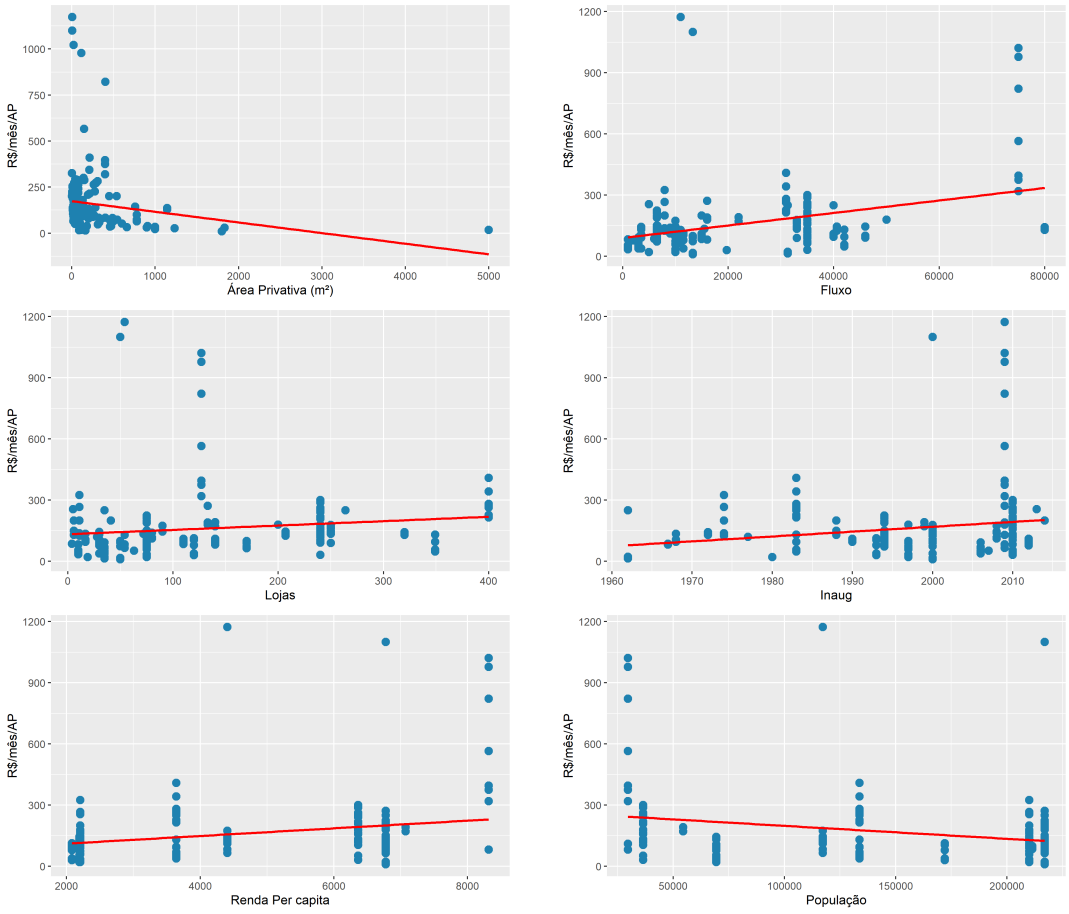

Figura 3: Gráficos de Dispersão entre as Variáveis Quantitativas e a Variável Dependente.

Observando os gráficos da Figura 3, com todos os dados da amostra, podemos

notar que para as variáveis de Área Privativa e População, há uma associação, embora fraca, negativa em relação ao valor unitário locativo. Essa primeira relação atende as expectativas iniciais pois é de conhecimento que o valor unitário é inversamente proporcional a área de um imóvel, quanto para valores locativos quanto para valores venais. Já a segunda relação, entre o número de população da região e o valor unitário, não é possível tirar conclusões claras.

Para as variáveis Fluxo, Lojas, Inauguração e Renda Per Capita, a relação aparenta ser, embora fraca, positiva. Para essas variáveis, esse tipo de relação em uma análise simples, parece satisfazer as expectativas, pois, um empreendimento com um fluxo maior, será mais atrativo para comerciantes e consequentemente poderá dispor de um espaço com uma área maior, abrigando assim uma quantidade maior de lojas. Pensando na variável Inauguração, um empreendimento que seja atuante por mais tempo no mercado, tende a conseguir formar um público alvo considerável que venha a utilizar daquele espaço com maior frequência. Por fim, uma região que contenha uma Renda Per Capita maior trará ao empreendimento e consequentemente a loja, um público que consequentemente poderá influenciar diretamente na receita final.

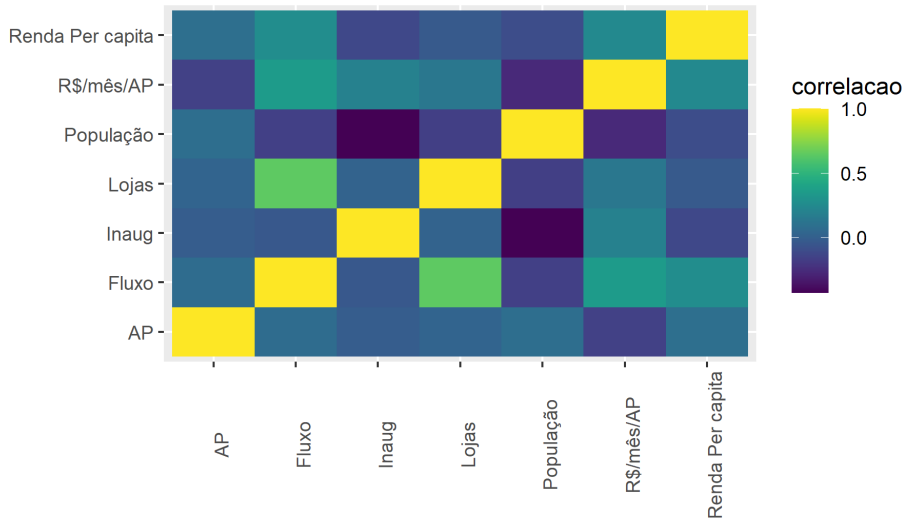

Figura 4: Gráfico de Calor para Correlação entre as Variáveis Qualitatias e a Variável Dependente.

O gráfico de calor, Figura 4, nos mostra as correlações entre todos os pares de variáveis quantitativas em que as cores mais próximas ao tom do amarelo significam uma associação forte positiva, valores próximos a 1, entre as variáveis, as cores mais próximas ao verde demonstram uma não associação entre esses pares, valores próximos a 0 e as cores mais próximas ao azul escuro uma associação forte negativa, valores próximos a -1.

Para essa análise inicial é necessário um ponto de atenção pois, como descrito na Norma (NBR-14.653-1, 2019), uma alta correlação entre duas variáveis independentes

pode provocar degeneração no modelo resultando em variâncias muito grandes das estimativas dos parâmetros e consequentemente acarretarem na aceitação da hipótese nula e a possível eliminação de variáveis fundamentais para compor o modelo.

Analisando primeiramente a correlação entre a variável dependente R$/mês/AP e as demais variáveis independentes, vemos que como observado na análise dos gráficos de dispersão, a variável AP e População apresenta uma correlação fraca moderada, negativa e a uma correlação embora moderada, positiva, com as variáveis Fluxo e Renda per capita.

Entre os pares de variáveis independentes, é necessário a atenção entre as variáveis Lojas e Fluxo, com relação positiva e entre as variáveis População e Inauguração, com relação negativa. A primeira relação por sua vez, como descrito na análise acima dos gráficos de dispersão, pode ser explicada, porém a segunda, não há uma análise de associação simples e direta que possa se fazer.

Variáveis Qualitativas Nominais

Para a análise das variáveis qualitativas, foram criados *BoxPlots* separados por categoria da variável em relação ao seu valor unitário locativo, para um primeiro estudo visual da mesma. Optou-se por separar a análise entre as variáveis qualitativas nominais, usualmente chamadas de variáveis dicotômicas no meio da Engenharia de Avaliação e as variáveis qualitativas ordinais, usualmente chamadas de variáveis com códigos alocados. A ideia dessa divisão é que seja possível, assim como na análise das variáveis numéricas, analisar a tendencia linear dessas variáveis.

Devido a existência da variável Tipo de Empreendimento, não foi gerado os *Boxplots* das variáveis representativas do imóvel estar localizado ou não em uma Universidade, Mall/Posto e Shopping pois as mesmas análises serão feitas utilizando o gráfico dispostos na análise abaixo das variáveis qualitativas ordinais.

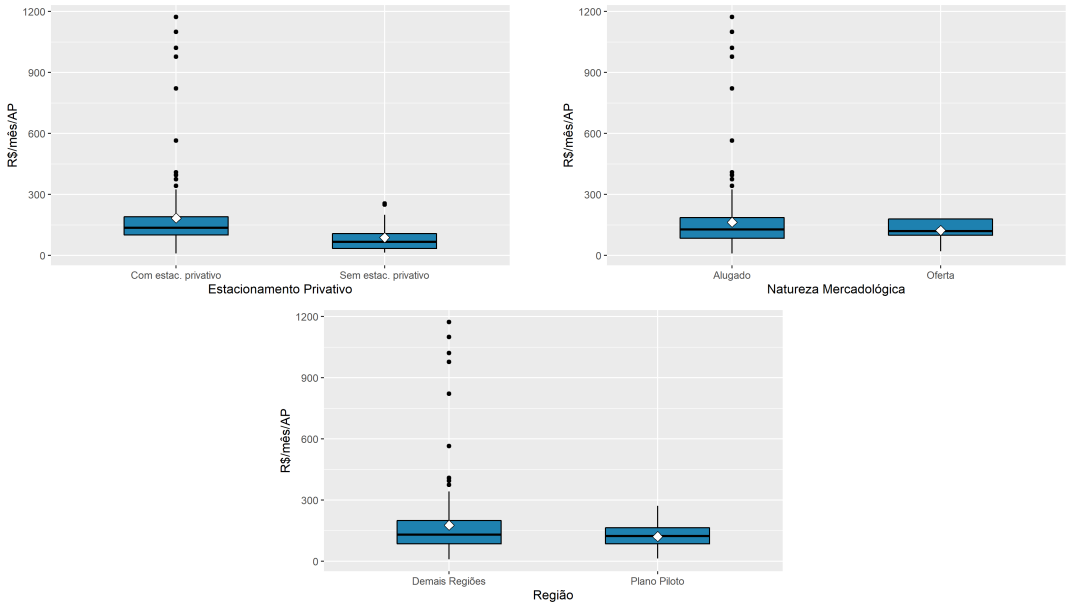

Figura 5: *BoxPlot* entre as Variáveis Qualitativas Nominais e a Variável Dependente.

Observando os gráficos da Figura 5, é possível notar que há visualmente uma certa igualdade na distribuição de ambas as categorias das variáveis Natureza Mercadológica e Região, porém, com muitos dados discrepantes para as categorias de imóveis já alugados e imóveis em regiões não sendo no Plano Piloto. É possível notar que estes dados realmente causam influencia pois o valor médio unitário locativo acaba se distanciando do valor da mediana. Outro ponto importante e observado juntamente com a observação do banco de dados é que há pouquíssimos imóveis nessa amostra que encontravam-se em situação de oferta, no momento da coleta dos dados. Visualmente, essas variáveis também não aparentam traze uma relação linear significante com a variável dependente.

Especificamente em relação ao gráfico da variável que descreve se o imóvel há ou não estacionamento privativo, conseguimos notar que a distribuição dos dados já aparenta, visualmente, uma certa diferença em que o valor unitário locativo dos imóveis com a presença de estacionamento privativo são ligeiramente superiores ao valor unitário locativos dos imóveis sem a presença de estacionamento privativo. Entretanto é importante observar que há uma quantidade com certa relevância de dados discrepância, assim como nos demais gráficos analisados.

Variáveis Qualitativas Ordinais

Para a análise das variáveis qualitativas ordinais, analogamente a análise anterior, serão expostos *Boxplots* porém, com a diferença de que as categorias das variáveis serão ordenadas de maneira que seja possível observar se há uma tendencia na distribuição dos seus valores entre estas categorias e o valor unitário locativo.

Para a variável segmento, serão utilizados os códigos numéricos de 1 a 5 para as

suas categorias devido ao tamanho da sua descrição que é, respectivamente a seguinte: Salas/Quiosque/Âncora, Diversões/Academia/Farmácia, Café/Bomboniere/Moda/Perfumaria, Restaurante/Lanchonete e por fim, Bancos/Agências.

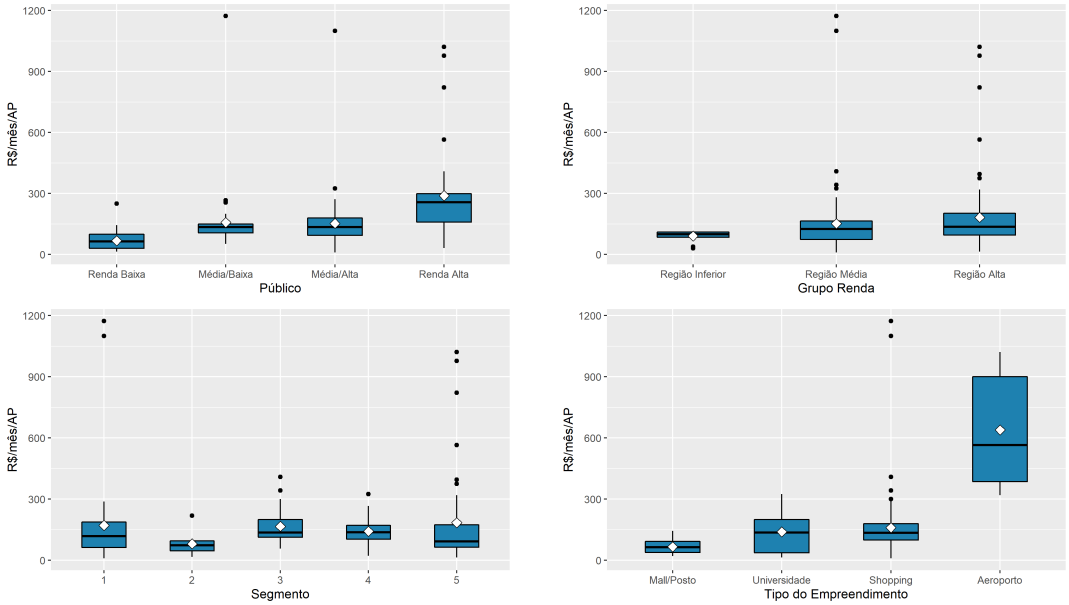

Figura 6: *BoxPlot* entre as Variáveis Qualitativas Ordinais e a Variável Dependente.

Inicialmente, para as quatro variáveis estudadas, observasse uma tendencia na sua distribuição de crescimento em que os valores unitários locativos das categorias iniciais são inferiores aos valores unitários locativos para as categorias finais, fato este que contempla as expectativas devido a forma como foram criadas e separadas as categorias. Observando o primeiro gráfico da Figura 6, é possível notar que embora essa relação da distribuição positiva aconteça, há pontos discrepantes muito acima dos valores padrões observados para as categorias do público com renda média/baixa e média/alta, que chegam a ter valores superiores até aos valores discrepantes que existem para o público com renda alta.

Na segunda situação, para a variável referente a divisão do grupo de renda, vemos que a distribuição dos dados para a região descriminada como inferior é bem menor do que para as demais, mas que ainda, observando ao todo as médias e medianas permanecem muito próximas para as três categorias. É possível notar também, assim como para a variável anterior, a presença de bastante dados discrepantes.

Pra a terceira situação, diferentemente das demais variáveis, a relação entre as suas distribuições não fica tão clara onde temos que o segundo código apresenta uma distribuição, abaixo dos imóveis com o segmento de Sala/Quiosque/Âncora, mas que pode ter sido altamente influenciado por dois valores discrepantes existentes neste código. Essa mesma diferença se dá também para o último e penúltimo código em relação as

lojas com o segmento de Café/Bomboniere/Moda/Perfume onde as demais apresentam distribuição dos seus valores inferior a este terceiro código. Novamente, há a presença de valores discrepantes e que para o segmento de lojas sendo Banco/Agência trouxe uma clara variância distanciando muito o valor da média para a mediana, nessa amostra.

Variável Temporal

Devido a importância e sensibilidade que tem em se tratar de uma variável temporal para situações econômicas de mercado como essa, ainda mais devido a diversas situações as quais o mundo passou nestes últimos anos, foi optado por se analisar separadamente a variável temporal Ano. Abaixo na Figura 7, está disposto um gráfico de dispersão entre a variável referente ao ano de coleta do dado e também a variável escolhida como dependente, R$/mês/AP.

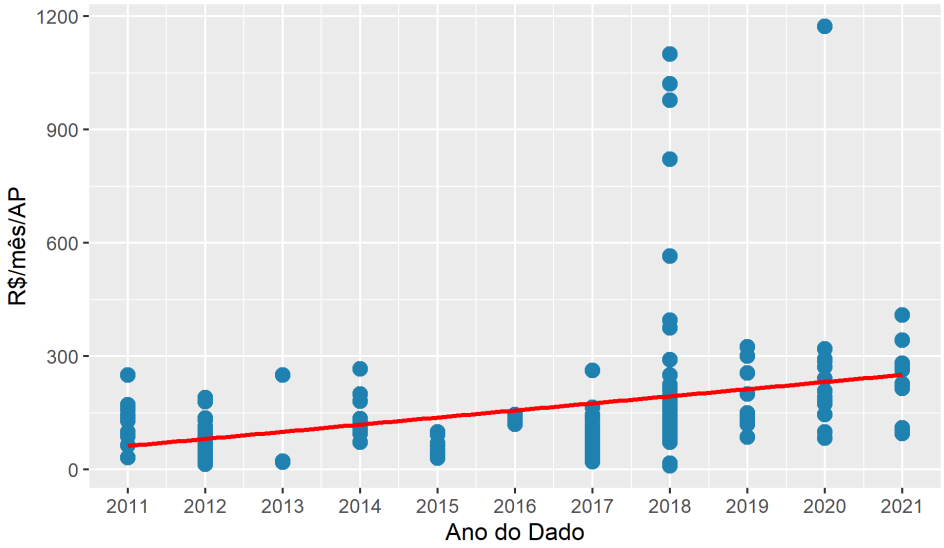

Figura 7: Gráfico de dispersão entre a Variável Ano e a Variável Dependente.

Em primeira análise, vale focar nos anos de 2020 e 2021 para observar se de alguma forma os dados obtidos neste ano trouxeram alguma mudança significativa devido a situação ao qual o país passou durante esses dois anos com a pandemia. Neste caso, observa-se que aparentemente, pelo menos para os dados contidos nessa amostra, não houve nenhuma alteração. Vemos também que a distribuição crescente de forma positiva parece existir e que no ano de 2018, há a presença de dados com valores discrepantes muito acima dos valores médios observados para aquele ano. Embora a variável exija um pouco mais de atenção, para essa amostra, o valor unitário locativo de lojas no decorre dos anos entre 2011 e 2021 se distribuíram de maneira normal.

Por fim, realizado o teste de correlação de Pearson entra a variável Ano e R$/mês/AP, vimos que o que foi observado na análise acima se concretiza. Com um resultado de 0,35, podemos afimar que há uma correlação positiva moderada, ou seja, os valores unitários

locativos tendem a crescer com o passar dos anos.

Análises

No geral, foi observado que as variáveis apresentaram em sua maioria, comportamento em relação as suas distribuições, esperados e uma boa distribuição dos seus dados quando comparados com variável determinante R$/mês/AP. Ficou notável também que há a presença de dados discrepantes que embora não tenham trago à primeira vista uma grande alteração nas análises podem ser prejudicais para um modelo futuro que virá a ser criado e trazendo de alguma forma uma alteração para uma futura projeção de valores locativos. Devido a isso, com o auxílio do Software SISDEA, que nos permite de maneira fácil visualizar quais desses dados tem sido discrepante, optou-se então por retirar da amostra para a realização das análises um total de sete dados.

Entre esses dados temos um que apresentou uma das menores área privativa do modelo, sendo de apenas 5,37m^2 e que consta com um valor altíssimo de 6.300 R$/mês o que explica pontualmente a sua discrepância de acordo com o restante da amostra. Foi observado também que dentre esses dados retirados, dois eram lojas localizados no Aeroporto Internacional de Brasília, situação essa peculiar de economia onde se sabe que os preços, não só de alugueis, nestes locais costumam ser superiores as demais situações. Por fim, observou-se também que haviam outros dois dados localizados em Clubes, o que pode vir a justificar ter tido esse destaque devido a alguma situação ao qual as variáveis utilizadas não puderam compreender essa diferença de valores.

Afim de encontrarmos o melhor modelo que possa predizer o valor locativo de uma loja, iremos aplicar técnicas de seleção de variáveis em modelos de regressão múltipla e também técnicas não paramétricas, ambas apoiadas do aprendizado de maquinas, para que assim, com o auxílio do SISDEA, seja possível a criação do melhor modelo possível e que contemple os parâmetros exigidos pela Norma (NBR-14.653-1, 2019).

Será utilizada a transformação de $Ln(x)$ da variável dicotômica R$/mês/AP afim de se reduzir o efeito do viés nos dados e tornar a comparação entre todos os empreendimentos constantes na amostra, que podem ser economicamente diferentes, mais justa. Também, já serão desconsiderados os dados encontrados como altamente discrepantes na análise anterior.

4.2 Seleção de Variáveis

Inicialmente foram retiradas as 6 primeiras variáveis do banco de dados, essas utilizadas para complemento e descrição das amostras coletadas, foi optado também pela utilização da variável R$/mês/AP como variável dependente e com isso, será retirada também das análises a variável referente ao valor total locativo R$/mês.

Como descrito e observado na seção anterior, trabalharemos então com um banco de dados com um total de 182 amostras e distribuídas em 18 variáveis e com a transformação de $Ln(x)$ aplicada a variável dependente.

Modelo completo

Inicialmente, ajustou-se o modelo completo contendo todas as 18 variáveis do banco de dados. Na Tabela 1 abaixo, temos descritos os principais resultados para esse modelo.

Tabela 1: Principais Resultados dos Coeficientes para o Modelo Completo

Variável	Estimativa	Erro Padrão	Estatística T	P-Valor
AP	<-0,01	<0,01	-5,11	<0,01
Estac	0,83	0,25	3,31	<0,01
Fluxo	<-0,01	<0,01	-1,23	0,22
Lojas	<0,01	<0,01	2,94	<0,01
Inaug	0,02	<0,01	4,69	<0,01
Publico	0,07	0,09	0,74	0,46
Renda	<-0,01	<0,01	-0,14	0,89
GrupoRenda	-0,04	0,16	-0,26	0,79
População	<0,01	<0,01	0,45	0,65
Plan	0,22	0,20	1,08	0,28
Segmento	0,01	0,02	0,30	0,77
Shopp	-0,12	0,20	-0,59	0,55
MallPosto	3,05	0,89	3,42	<0,01
Univ	2,21	0,59	3,77	<0,01
Tipo	1,30	0,43	3,05	<0,01
Ano	0,06	0,01	4,71	<0,01
Ofer	-0,02	0,17	-0.14	0.90

Observado a Tabela 1 e adotando um nível de significância de $\alpha = 0,05$ é possível notar que temos 8 variáveis que são significativas, sendo elas: AP, Estac, Lojas, Inaug, Mall/Posto, Univ, Tipo e Ano. As demais variáveis apresentaram P-valores superiores a 0,22, valor esse referente a variável Fluxo e, portanto, não são significativas para essa primeira análise.

Tabela 2: Principais Estatísticas Para o Modelo Completo

Estatística	Valor
Erro Padrão	0,43
G.L.	164
R^2	0,65
$R^2_{ajustado}$	0,62
Estatística F	18,23
P-Valor	<0,01

De acordo com a Tabela 2, referente aos principais resultados estatísticos para o modelo de regressão completo, temos que o erro padrão estimado do modelo, ou seja, a estimativa de α possui um valor de 0,43 com 165 graus de liberdade. Temos ainda que o R^2 e o $R^2_{ajustado}$ são respectivamente iguais a 0,61 e 0,58. Com isso, podemos concluir que este modelo é capaz de explicar 61% da variável resposta para o caso R^2 e ajustado ao número de variáveis explicativas, o modelo completo é capaz de explicar 62% da variável resposta. Por fim, observando o P-valor encontrado para o modelo e considerando o nível de confiança de $\alpha = 0,05$, conclui-se que há regressão.

Ajuste com Pacote Leaps

Para que possamos encontrar a melhor combinação de variáveis possíveis para este modelo, utilizaremos a aplicação do pacote leaps, listando os k melhores modelos, que para cada número de variáveis encontrara o modelo segundo o critério de menor soma de quadrados residual, para conjuntos e preditores de um até 18 variáveis explicativas. Como isso, poderemos então plotar um gráfico com os valores de RSQ, RSS, AJDR2, CP e BIC, medidas essas que serão utilizadas como critérios de seleção destes modelos.

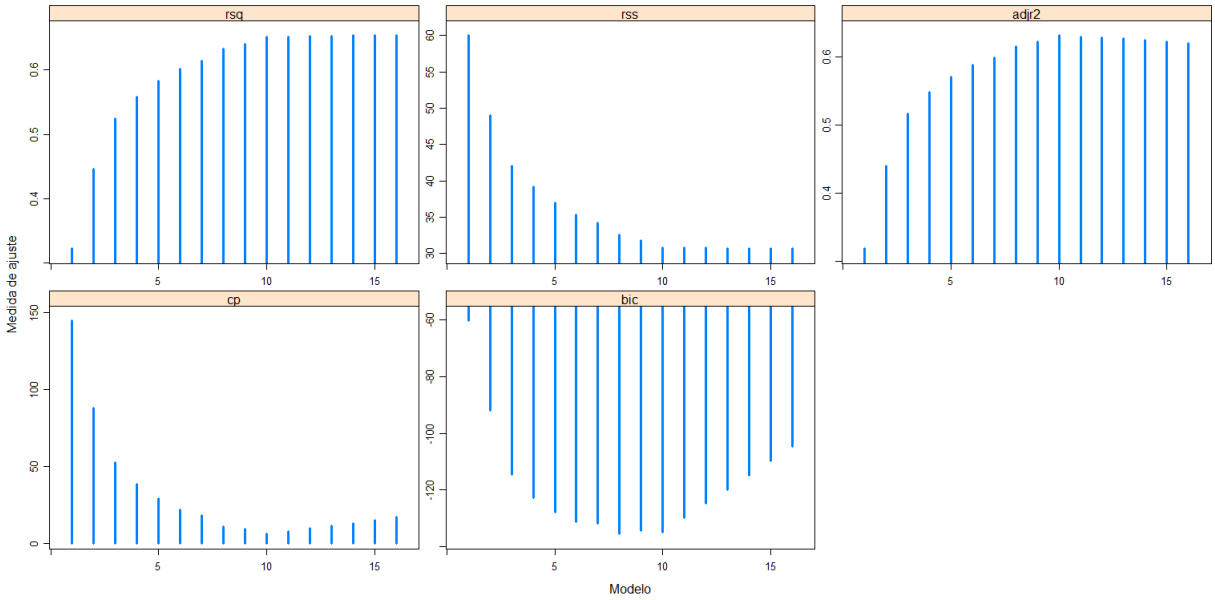

Figura 8: Gráfico das principais medidas de ajuste do modelo completo.

Observado o gráfico 8 referente as principais medidas de ajuste do modelo completo com as 18 variáveis obtido através do uso do pacote leaps no R, é possível observar que o modelo que melhor se adequa é o que contem 10 variáveis, tendo então um acréscimo de duas variáveis referente aos primeiros resultados obtidos na Tabela 1. Com isso, teremos então o modelo completo com as seguintes variáveis:

$$\begin{aligned}\ln R\$/mês/AP = \beta_0 + \beta_1 AP_i + \beta_2 Estac_i + \beta_3 Fluxo_i + \beta_4 Lojas_i + \beta_5 Inaug_i \\ + \beta_6 Plan_i + \beta_7 MallPosto_i + \beta_8 Univ_i + \beta_9 Tipo_i + \beta_{10} Ano_i\end{aligned}$$

Abaixo na Tabela 3, estão dispostos os principais resultados para o modelo completo ajustado com as dez variáveis escolhidas.

Tabela 3: Principais Resultados dos Coeficientes para o Modelo Completo Ajustado

Variável	Estimativa	Erro Padrão	Estatística T	P-Valor
AP	<-0,01	<0,01	-5,72	<0,01
Estac	0,94	0,19	4,99	<0,01
Fluxo	<-0,01	<0,01	-2,32	0,02
Lojas	<0,01	<0,01	5,66	<0,01
Inaug	0,02	<0,01	5,89	<0,01
Plan	0,22	0,09	2,37	0,02
MallPosto	3,34	0,69	4,84	<0,01
Univ	2,51	0,40	6,27	<0,01
Tipo	1,39	0,33	4,25	<0,01
Ano	0,06	0,01	5,79	<0,01

Como observado na Tabela 3, todas as variáveis são significativas, adotando um $\alpha = 0,05$ e diferentemente dos primeiros resultados obtidos na Tabela 1 agora temos a presença de duas novas variáveis, Fluxo e Plan.

Tabela 4: Principais Estatísticas Para o Modelo Completo Ajustado

Estatística	Valor
Erro Padrão	0,42
G.L.	171
R^2	0,65
$R^2_{ajustado}$	0,63
Estatística F	31,99
P-Valor	<0,01

Acima, na Tabela 4, temos as principais estatísticas para o modelo completo ajustado com dez variáveis e que se comparado com os resultados do modelo completo inicial, disposto na Tabela 2 temos um aumento nos graus de liberdade de 164 para 171, na estatística F de 18,23 anteriormente para 31,99 e uma pequena melhora para o $R^2_{ajustado}$.

Modelo Alternativo

Como observado no banco de dados, temos três variáveis dicotômicas, sendo elas Shopp, Mall/Posto e Univ que de certa são agregadas em uma única variável qualitativa ordinal, variável Tipo. Devido a isso, será então testado um modelo alternativo dispensando o uso dessas variáveis dicotômicas afim de se ter uma alternativa para esse estudo.

A análise será análoga a anterior feita para o modelo completo porém, com apenas 15 variáveis e os mesmos 182 dados permanecendo com a transformação de $Ln(x)$ da

variável dependente R$/mês/AP. Abaixo na Tabela 5, temos os principais resultados para o modelo alternativo.

Tabela 5: Principais Resultados dos Coeficientes para o Modelo Alternativo

Variável	Estimativa	Erro Padrão	Estatística T	P-Valor
AP	<-0,01	<0,01	-5,91	<0,01
Estac	0,2	0,21	0,95	0,34
Fluxo	<0,01	<0,01	2,15	0,03
Lojas	<-0,01	<0,01	-1,09	0,27
Inaug	<0,01	<0,01	2,73	<0,01
Publico	0,31	0,07	4,34	<0,01
Renda	<-0,01	<0,01	-0,92	0,36
GrupoRenda	-0,12	0,16	-0,76	0,44
População	<0,01	<0,01	0,38	0,71
Plan	0,13	0,21	0,62	0,54
Segmento	0,02	0,03	0,74	0,46
Tipo	<-0,01	0,14	-0,05	0,96
Ano	0,08	0,01	5,59	<0,01
Ofer	-0,31	0,16	-1,93	0,06

De acordo com a Tabela 5 e seguindo com o nível de significância adotado anteriormente de $\alpha = 0,05$, vemos que 5 variáveis mostraram-se significativas, sendo ela: AP, Fluxo, Inaug,Público e Ano. Se comparado coma a análise inicial sem ajuste do modelo completo, temos a presença de duas novas variáveis, Fluxo e Público, sendo a primeira contemplada por ajuste do modelo completo. As demais variáveis, de acordo com a 5, não foram significativas.

Tabela 6: Principais Estatísticas Para o Modelo Alternativo

Estatística	Valor
Erro Padrão	0,45
G.L.	167
R^2	0,61
$R^2_{ajustado}$	0,58
Estatística F	19
P-Valor	<0,01

A Tabela 6, referente aos principais resultados estatísticos para o modelo de regressão completo, nos demonstra que o erro padrão estimado do modelo, ou seja, a a

estimativa de α possui um valor de 0,45 com 167 graus de liberdade, superior ao modelo completo sem ajuste. Temos ainda que o R^2 e o $R^2_{ajustado}$ são respectivamente iguais a 0,61 e 0,58. Com isso, podemos concluir que este modelo é capaz de explicar 61% da variável resposta para o caso R^2 igualmente ao modelo anterior, porém, para o caso ajustado ao número de variáveis explicativas esse valor é inferior aos modelos anteriores, sendo de 58%. Por fim, o P-valor encontrado para este modelo e seguinte os padrões de adoção do nível de confiança, conclui-se que há regressão.

Ajuste com pacote Leaps

Seguindo os padrões da análise, utilizaremos novamente o pacote do R leaps porém, para um conjunto de preditores de um até 15 variáveis axplicativas. Com isso, poderemos então plotar um gráfico com os valores de RSQ, RSS, ADJR2, CP e BIC, medidas estás que serão utilizadas como critérios de seleção destes modelos.

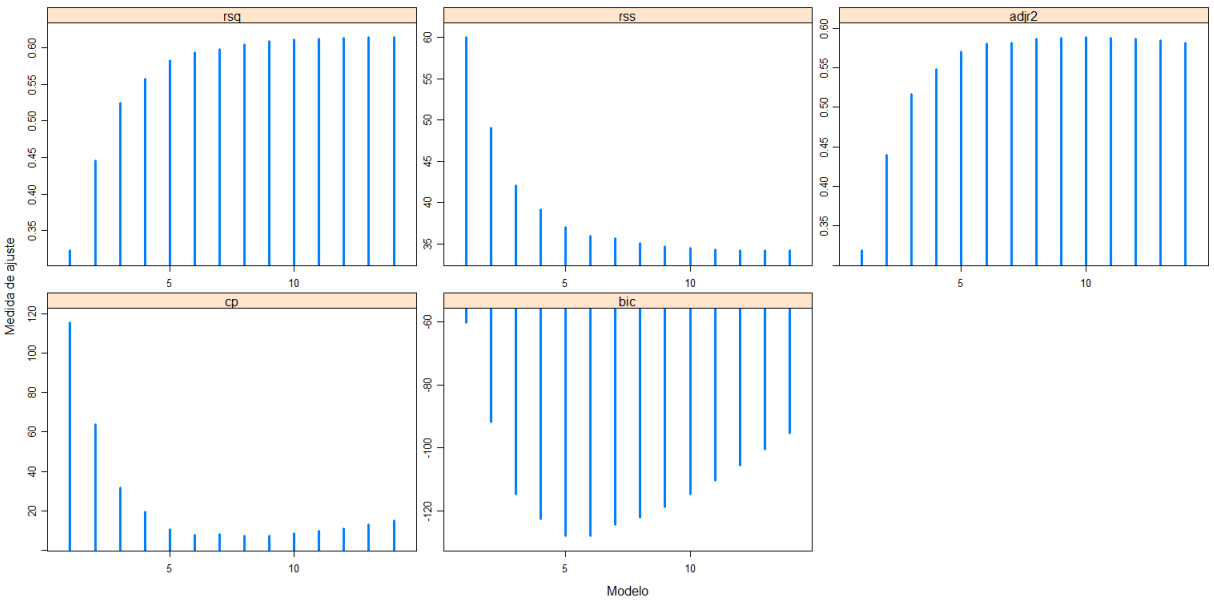

Figura 9: Gráfico das principais medidas de ajuste do modelo alternativo.

Observado o gráfico 9 referente as principais medidas de ajuste do modelo alternativo com as 15 variáveis obtido através do uso do pacote leaps no R, é possível observar que o modelo que melhor se adequa é o que contem 5 variáveis, seguindo o que se foi observado na Tabela 5. Com isso, teremos então o modelo alternativo com as seguintes variáveis:

$$\ln R\$/mês/AP = \beta_0 + \beta_1 AP_i + \beta_2 Publico_i + \beta_3 Renda_i + \beta_4 Tipo_i + \beta_5 Ano_i$$

Abaixo na Tabela 7, estão dispostos os principais resultados para o modelo alter-

nativo ajustado com as cinco variáveis escolhidas.

Tabela 7: Principais Resultados dos Coeficientes para o Modelo Alternativo Ajustado

Variável	Estimativa	Erro Padrão	Estatística T	P-Valor
AP	<-0,01	<0,01	-6,14	<0,01
Inaug	<0,01	<0,01	2,22	0,03
Publico	0,28	0,05	5,82	<0,01
Renda	<-0,01	<0,01	-3,23	<0,01
Tipo	0,19	0,06	3,25	<0,01
Ano	0,06	0,01	5,18	<0,01

Como observado na Tabela 7, todas as variáveis são significativas, adotando um $\alpha = 0,05$ igualmente as análises que se seguiam este modelo alternativo.

Tabela 8: Principais Estatísticas Para o Modelo Alternativo Ajustado

Estatística	Valor
Erro Padrão	0,45
G.L.	175
R^2	0,59
$R^2_{ajustado}$	0,58
Estatística F	42,7
P-Valor	<0,01

Por fim, observado a Tabela 8 referente as principais medidas do modelo alternativo ajustado e comparando com as mesmas medidas para o modelo completo ajustado, dispostos na Tabela 8, vemos que embora a Estatística F, o Erro Padrão e os graus de liberdade tenham sidos superiores, os valores de R^2 e $R^2_{ajustado}$ foram ambos inferiores , ou seja, a capacidade de explicação do modelo alternativo ajustado, para ambas as opções é inferior a capacidade de explicação do modelo completo ajustado.

4.3 Validação Cruzada K-Fold

Será utilizado a validação cruzada para selecionar o número ideal de variáveis para a predição, dessa forma, a função custo será o erro quadrático médio (MSE) da predição do conjunto de teste. Para isso, será utilizado a validação cruzada 5-fold com 10 repetições independentes. Vale ressaltar que em cada fold que será criado pelas partições, o conjunto das variáveis testadas não serão as mesmas. Por fim, após a determinação do número de variáveis ideais, será determinação quais serão essas variáveis.

Abaixo, observando o gráfico 10, referente a aplicação da validação cruzada para o modelo completo vemos que a distribuição para a função custo que esta sendo utilizada pelo erro quadrático médio, apresenta melhores resultados para um conjunto com dez variáveis, indo assim de acordo com as análises feitas na seção anterior.

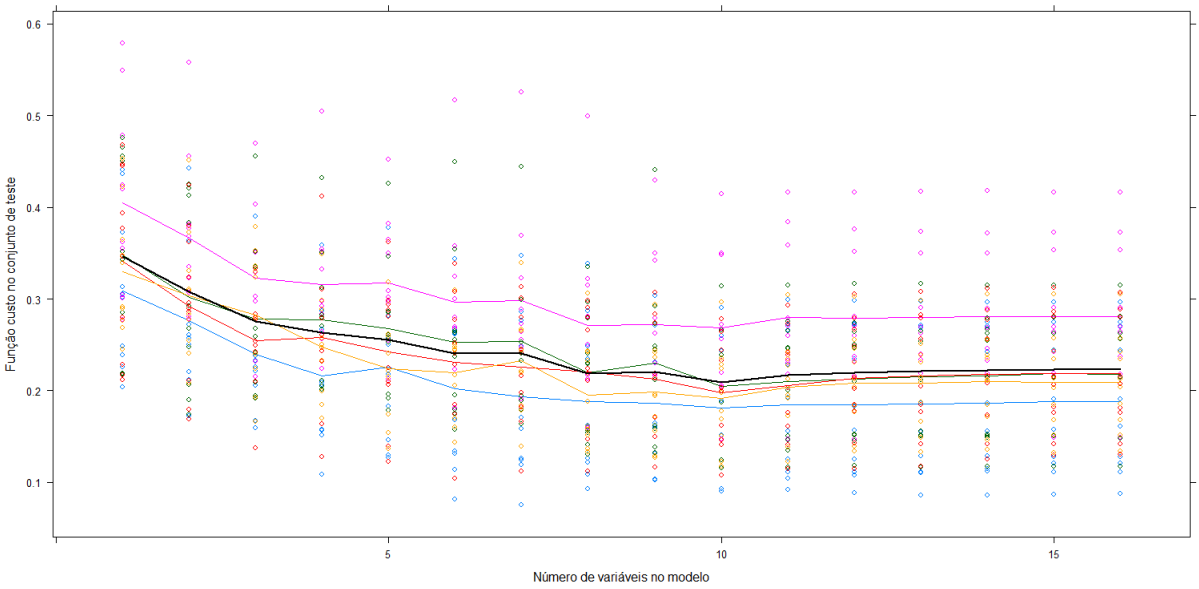

Figura 10: Gráfico da Função custo no conjunto de teste pelo número de variáveis no modelo completo.

Utilizando então a função regsubsets no R, determinamos as melhores variáveis, dispostas abaix,

$$\begin{aligned}\ln R\$/mês/AP = \beta_0 + \beta_1 AP_i + \beta_2 Estac_i + \beta_3 Fluxo_i + \beta_4 Lojas_i + \beta_5 Inaug_i \\ + \beta_6 Plan_i + \beta_7 MallPosto_i + \beta_8 Univ_i + \beta_9 Tipo_i + \beta_1 0 Ano_i\end{aligned}$$

Resultado este, de acordo com o que se segue a análise.

Analogamente, observando o Figura 11, referente a aplicação da validação cruzada para o modelo alternativo, vemos que a distribuição para a função custo que está sendo utilizada de acordo com o erro quadrático médio, apresentou melhores resultados para o conjunto com 5 variáveis.

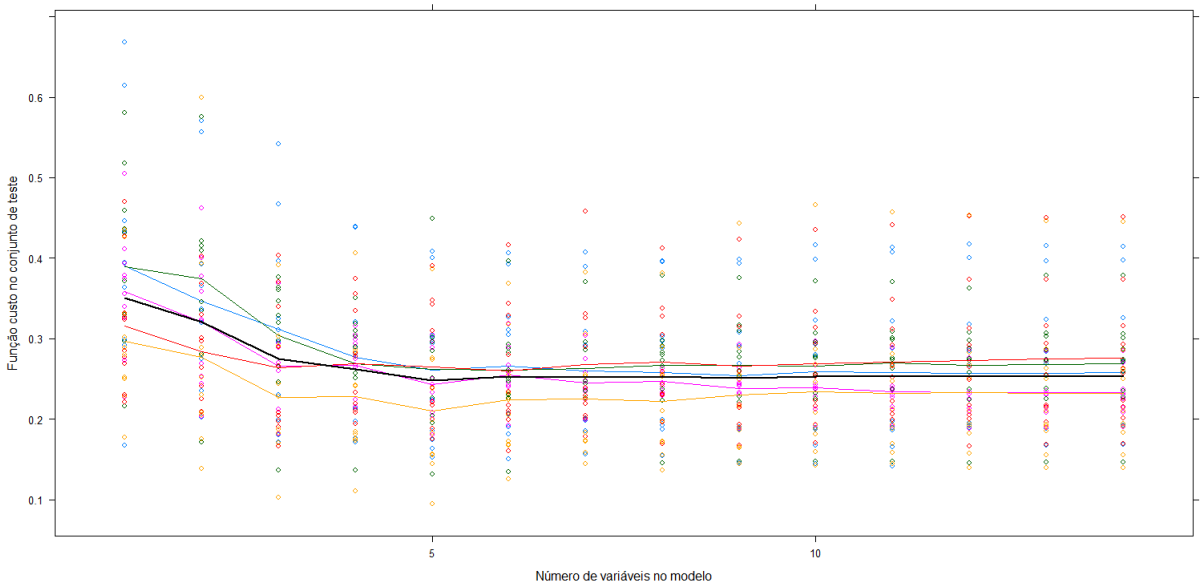

Figura 11: Gráfico da Função custo no conjunto de teste pelo número de variáveis no modelo alternativo.

Aplicando a função regsubsets no R, determinamos as melhores variáveis, descritas abaixo.

$$\ln R\$/mês/AP = \beta_0 + \beta_1 AP_i + \beta_2 Publico_i + \beta_3 Renda_i + \beta_4 Tipo_i + \beta_5 Ano_i$$

Resultado este, assim como para o modelo completo ajustado, igual as análises que se seguem para o modelo alternativo ajustado.

4.4 Árvore de Regressão e Florestas Aleatórias

A partir da criação de uma separação de um conjunto de treino e teste, do banco de dados que esta sendo utilizado, com 182 dados e 18 variáveis, onde será sortido aleatoriamente 70% das amostras para compor o conjunto de treino e os restantes 30% para compor o conjunto de teste, será criado uma árvore de regressão utilizando a função rpart no R.

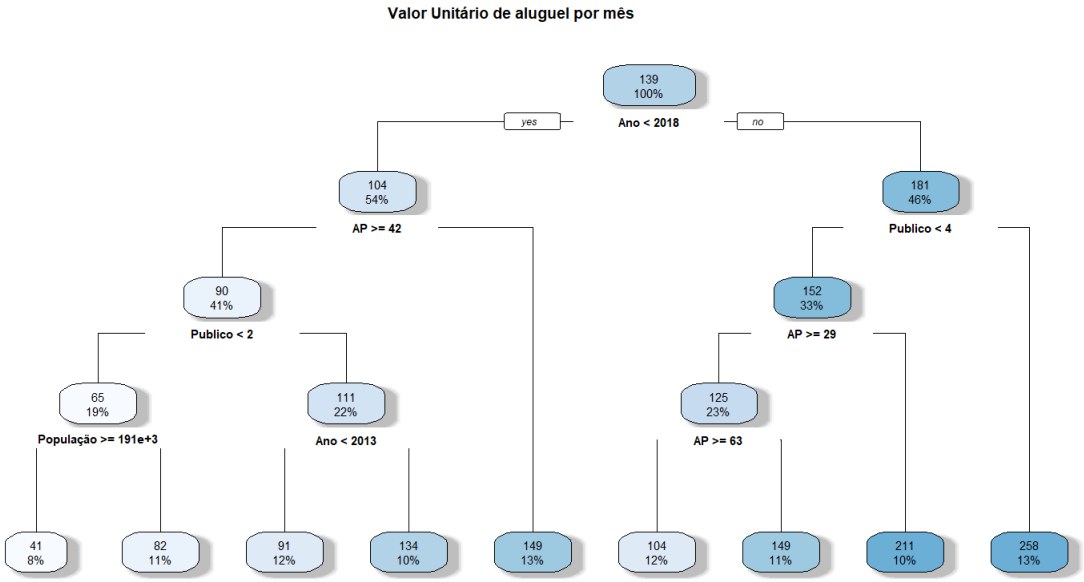

Figura 12: Estrutura da árvore de regressão para o valor locativo do imóvel.

A Figura 12, construída por particionamentos denominados de nó e cada resultado a partir deste nó chamados de folha, nos permite analisar a relação entre variáveis explicativas e a variável resposta. O tamanho de cada ramo desta árvore é definido proporcionalmente à diminuição do erro quadrático médio que ocorreu quando a respectiva foi criada.

Para o caso observado, temos a presença de quatro variáveis explicativas sendo elas: Ano, Publico,AP e População. A partir do primeiro nó, podemos notar que os dados com anos inferiores a 2018 tem uma queda no seu valor unitário de locação de 139 para 104 R$/mês/AP, sendo esses referentes a 54% da amostra estudada. Temos também, que os dados com ano superior a 2018 e público com código superior a 3, apresentam um valor unitário locativo de 258 R$/mês/AP, representando 13% da amostra estudada. Para a maior repartição, representante de 8% dos dados estudados, temos os imóveis com ano de coleta inferior a 2018, área privativa inferior a 42m^2, público inferior ao código alocado 2 e população inferior a 191.000, que resulta em um valor de 41 R$/mês/m^2.

Por fim, utilizaremos o algoritmo da árvore de decisão como base para um modelo mais avançado, Random Forest. Será criado um número N de árvores a partir de uma amostragem de um número $p2 < p1$ de colunas, onde $p1$ é a quantidade total de colunas no conjunto de dados. Para a visualização desses resultados, utilizaremos o cálculo da importância de cada variável, medida através do calculo do total da redução da soma de quadrados dos resíduos para todas as divisões que foram feitas com base nessa variável. Esse processo será repetido para todas as árvores e por fim será calculado a média dessa redução para ser utilizada como medida de importância para cada variável.

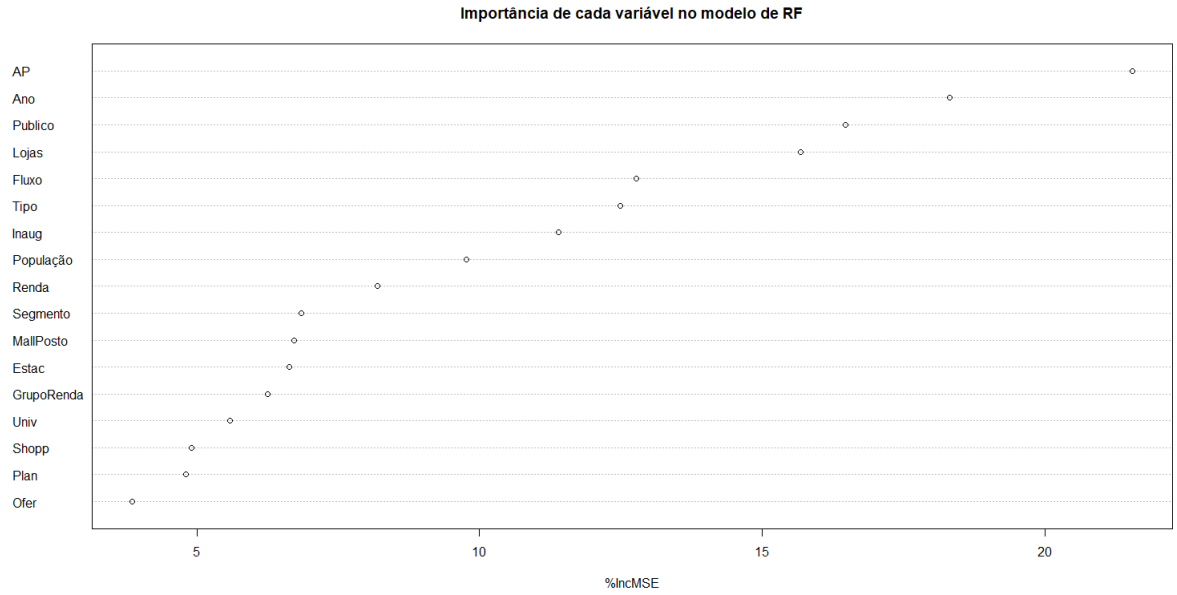

Figura 13: Importância de cada covariável do conjunto de dados na predição da variável R$/mês/AP.

4.5 Modelo Final - SISDEA

De acordo com as análises realizadas e os resultados encontrados, foi utilizado o Software SISDEA para a criação e ajuste do modelo de regressão linear múltipla, usualmente utilizado na Engenharia de Avaliação, com foco nas principais variáveis apontadas nas análises anteriores. Com isso, foi construído o modelo final cuja equação foi:

$$\ln(R\$/mês/AP) = -89,1549 - 0,2421 * ln(AP) + 0,2066 * Estac = 2 + 0,0019 * Lojas + 0,0050 * Inauguração + 0,1779 * Público + 0,0416 * Ano \quad (4.5.1)$$

Observando a equação 4.5.1, vemos que o modelo final foi construído com seis variáveis independentes sendo elas, AP, Estac=2, Lojas, Inauguração, Público e Ano. Além disso, dos 182 dados utilizadas foram considerados 149, sendo os demais desconsiderado através das análises gráficas, de resíduo e também de aderência, possibilitadas pelo Software. Na Tabela 9, estão expostas os principais resultados do modelo:

Tabela 9: Principais Estatísticas do Modelo Final em SISDEA

Estatística	Valor
Erro Padrão	0,22
G.L.	148
R^2	0,84
$R^2_{ajustado}$	0,81
Estatística F	124,46
P-Valor	<0,01

De acordo com a Tabela 9, referente aos principais resultados estatísticos para o modelo de regressão completo, temos que o erro padrão estimado do modelo, ou seja, a estimativa de α possui um valor de 0,22 com 148 graus de liberdade. Temos ainda que o R^2 e o $R^2_{ajustado}$ são respectivamente iguais a 0,84 e 0,81, que se comparados com os resultados para os modelos propostos nas Tabelas 2 e 4, nota-se uma grande melhora.

Com isso, podemos concluir que este modelo é capaz de explicar 84% da variável resposta para o caso R^2 e ajustado ao número de variáveis explicativas, o modelo completo é capaz de explicar 81% da variável resposta. Por fim, observando o P-valor encontrado para o modelo e considerando o nível de confiança de $\alpha = 0,05$, conclui-se que há regressão.

Análise dos Pressupostos do Modelo

Abaixo serão demonstrados, com resultados diretamente extraídos do modelo final em SISDEA a análise e o atendimento aos principais pressupostos do modelo de regressão linear múltipla.

Normalidade dos Resíduos

Através dos resultados obtidos pelo SISDEA é possível comparar as frequências relativas dos resíduos amostrais padronizadas nos intervalos [-1;+1],[-,1,64;+1,64] e [-1,96;+1,96], com as probabilidades da distribuição normal padrão nos mesmos intervalos, ou seja, 68%, 90% e 95%. Para o modelo construído, os respectivos valores encontrados foram de 68%, 87% e 94%, sendo assim aceitáveis para o atendimento a este pressuposto. Além disso e seguindo nesta análise, foi observado a Figura 14 referente aos resíduos amostrais padronizados no qual podemos notar uma simetria e um formato semelhante ao da curva da Normal.

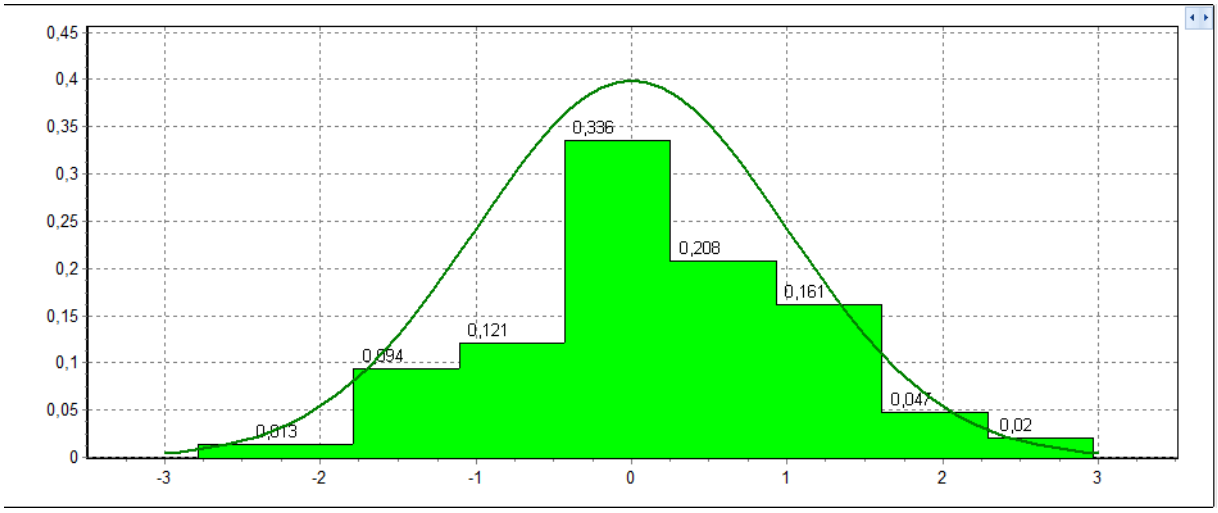

Figura 14: Histograma dos Resíduos Amostrais Padrozinados do Modelo em SISDEA.

Linearidade

A linearidade do modelo é a relação linear entre a variável dependente e as variáveis independentes utilizadas na formação da equação (4.5.1). Para essa análise, foi utilizado a Figura 15, onde podemos observar a relação entre as variáveis.Vale ressaltar que a variável dependente AP sofreu uma transformação para *Ln* forçando assim o atendimento a este ponto. Com isso o modelo demonstrou seguir o pressuposto de linearidade.

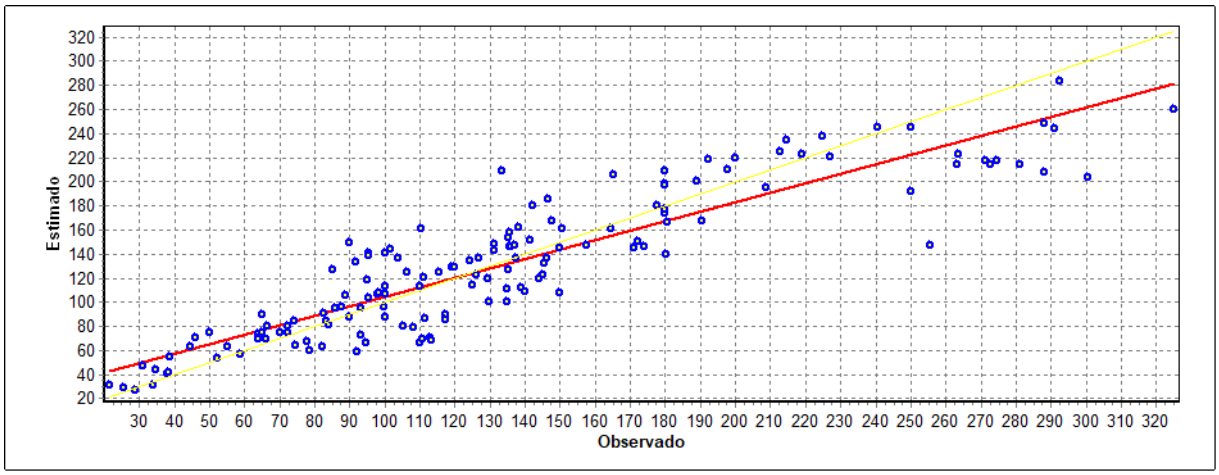

Figura 15: Gráfico de aderência entre os valores observados e estimados do Modelo em SISDEA.

Homocedasticidade

Observando a Figura 16, foi analisado a homoscedasticidade do modelo, ao qual podemos observar a forma aleatório como se da a distribuição dos pontos sem haver nenhum padrão definido dos mesmos em torno da reta horizontal que passa pela origem. Com isso, podemos concluir também que o critério de homoscedasticidade do modelo foi

respeitado.

Além disso, a partir deste gráfico, podemos observar também que há a presença de 7 dados denominados como Outliers, representado pelo ponto fora da faixa entre as retas horizontais, ou seja, um ponto que contém grande resíduo em relação aos demais que compõem a amostra, sendo representante de 4,70% dos dados da amostra estando assim abaixo dos 5% e dentro dos parâmetros permitidos pela Norma (NBR-14.653-2, 2011).

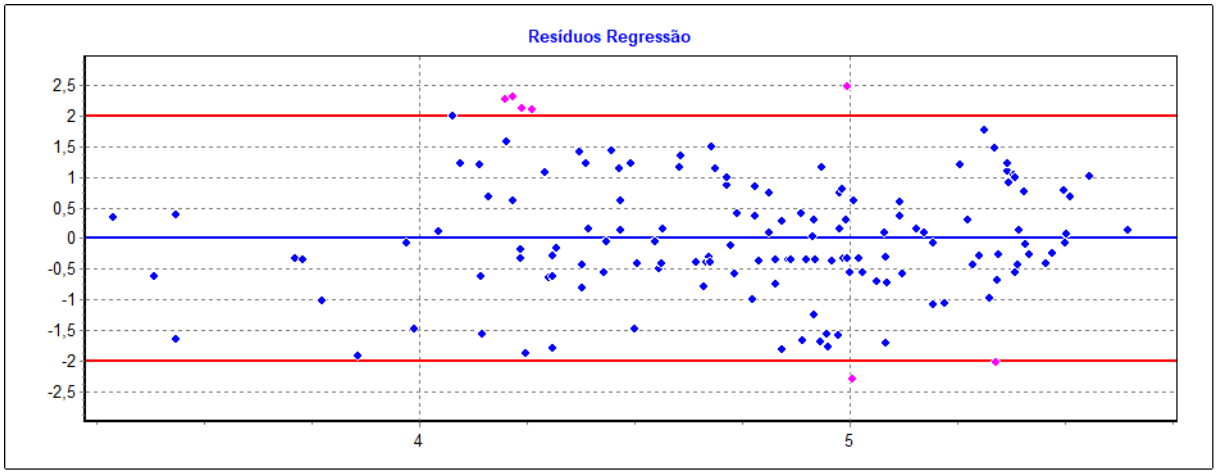

Figura 16: Gráfico dos Resíduos Relativos contra Resíduos Observados para o Modelo em SISDEA.

Multicolinearidade

Com o auxílio da Figura 17, referente a matriz de correlação entre os pares da variáveis utilizadas pelo modelo, nota-se que não houveram resultados superiores a 0,80 e devido a isso, o pressuposto de Multicolinearidade foram atendidos seguindo a recomendação da (NBR-14.653-2, 2011).

I..	Variável	Tran...	Alias	x1	x2	x3	x4	x5	x6	y
	AP	ln(x)	x1	0	0,00	0,26	0,11	-0,04	-0,17	-0,49
	Estac=2	x	x2	0,00	0	0,49	-0,15	0,63	0,11	0,53
	Lojas	x	x3	0,26	0,49	0	0,01	0,53	0,18	0,49
	Inauguração	x	x4	0,11	-0,15	0,01	0	-0,18	-0,05	-0,05
	Público	x	x5	-0,04	0,63	0,53	-0,18	0	0,32	0,68
	Ano	x	x6	-0,17	0,11	0,18	-0,05	0,32	0	0,50
	R$/mês/AP	ln(y)	y	-0,49	0,53	0,49	-0,05	0,68	0,50	0

Figura 17: Correlação entre os pares de variáveis para o Modelo em SISDEA.

Exemplo de uma Projeção

Para fator de teste do modelo e também para que seja possível o enquadramento

do mesmo dentro dos parâmetros definidos pela (NBR-14.653-2, 2011) para precisão e fundamentação, será realizado uma projeção do valor locativo de um imóvel com a seguintes características:

- AP = 250,00
- Estac=2 = 2
- Lojas = 250
- Inauguração = 2.011
- Público = 2
- Ano = 2.017

Por padrão, utilizando um Nível de confiança de 80%, foram obtidos os seguintes intervalos e valores para a projeção descrita acima:

Intervalo	Valor Unitário	Valor Total
Mínimo	(5,39%) 104,96	26.239,25
Médio	110,94	27.734,89
Máximo	(5,70%) 117,26	29.315,78

Tabela 10: Resultado da Projeção do Valor Locativo

Com isso, observando a Tabela 10, concluímos que um imóvel, diretamente comparativo com a amostra coletada e tratada, com as características mencionadas acima terá um valor médio locativo de aproximadamente R$ 27.000,00 e valor máximo e mínimo aproximados e respectivos de R$ 26.000,00 e R$ 29.000,00 a um nível de confiança de 80%.

Grau de Precisão e Fundamentação

A precisão de acordo com a (NBR-14.653-2, 2011) é estabelecida mediante análise da amplitude do intervalo de confiança de 80% em torno do valor central da estimativa. Observando a Figura 18 e de acordo com o intervalo obtido, para o modelo e projeção feita acima, de 11,09%, podemos concluir que o grau de precisão é 3.

Descrição	Graus de precisão		
	III	II	I
Amplitude do intervalo de confiança de 80% em torno do valor central da estimativa	≤30%	≤ 40%	≤ 50%

Figura 18: Grau de precisão em caso de utilização da regressão linear (imóveis urbanos e rurais).

Fonte: (NBR-14.653-2, 2011).

Por fim, para determinação da fundamentação, utiliza-se a pontuação obtida de acordo com a tabela mencionada na (NBR-14.653-2, 2011) e comparada com a tabela da Figura 19. Para o modelo construído, foi obtido 17 pontos para a tabela de fundamentação e por contemplar os critérios expostos na Figura 19, foi obtido o grau de fundamentação 3.

Graus	**III**	**II**	**I**
Pontos mínimos	**16**	**10**	**6**
Itens obrigatórios	2, 4, 5 e 6 no grau III e os demais no mínimo grau II	2, 4, 5 e 6 no mínimo grau II e os demais no mínimo grau I	Todos, no mínimo grau I

Figura 19: Grau de fundamentação em caso de utilização da regressão linear (imóveis urbanos e rurais).

Fonte: (NBR-14.653-2, 2011).

5 Conclusão

A metodologia utilizada na Engenharia de Avaliação, mais especificamente o método comparativo direto de dados de mercado é um processo já bem estruturado e dentro de um mercado imobiliária que conta com um grande desenvolvimento, aumentando assim ainda mais a complexidade, a aplicabilidade e o retorno dos estudos dentro desta área. Anterior a essa metodologia atual aplicada, este tipo de mercado era estudado através de cálculos simplificados e métodos rudimentares. Com os avanços tecnológicas e científicos, esse processo desenvolveu-se e possibilitou a aplicação de técnicas mais avançadas com grande destaque para inserção da inferência estatística que resultou em um aumento na qualidade e confiabilidade dos estudos e conclusões para o ramo imobiliário.

Como descrito pelo Engenheiro Pelli em (NETO, 2021) , a aplicação de novos métodos e a introdução de novas analises nesse tipo de mercado teve um avanço ainda mais crescente devido aos estudos das diversas variáveis que compõe o preço e consequentemente o valor dos bens realizados por profissionais autônomos, empresas públicas e privadas, entidades de classe e, principalmente, por profissionais do meio acadêmico, garantindo conclusões ainda mais seguras nos trabalhos avaliatórios.

O processo, comumente utilizado para que se chegue à conclusão do valor de um imóvel é focado principalmente nos fundamentos da inferência estatística e na Regressão Linear Múltipla, processo este em sua grande maioria realizado com o auxílio do SISDEA software para modelagem de dados. Seguindo a linearidade do constante avanço da Engenharia de Avaliação através de novas aplicações de metodologias neste processo, foi aplicado conhecimentos estatístico, desde a manipulação do banco de dados até o ajuste final do modelo em SISDEA, visando a melhora da criação de um modelo utilizado para a projeção do valor locativo de um imóvel e assim possibilitando uma melhor conclusão.

Inicialmente, na seção 3.2, foi reorganizado o banco de dados e adicionado novas variáveis e assim prosseguindo para a análise primaria da visualização da distribuição dos dados, por meio da análise descritiva, na seção 4.1. Posteriormente a isso, foi realizado o processo da seleção de variáveis com o auxilio de técnicas de aprendizado de maquinas, dividindo o banco de dados em duas partes sendo a de treino utilizada para o ajuste do modelo e para a testagem utilizado a segunda parte do banco, sendo a de teste, por meio da técnica K-Fold.

Foi escolhido trabalhar primeiramente com o modelo de regressão linear no software R, já introduzido por meio do software SISDEA dentro da engenharia de avaliação, construindo então dois modelos com conjuntos de variáveis diferentes. O uso dessa técnica nos permitiu ter uma análise mais completa entre as relações das variáveis independentes

com a variável dependente, bem como o comportamento que cada uma tinha em relação ao modelo. Utilizado a técnica paramétrica da regressão linear, optou-se pela aplicação também da árvore de regressão e Florestas Aleatórias, técnicas essas não paramétricas que embora não nos permita uma análise como a anterior, nos deu a possibilidade de analisar padrões nos dados que nos leva a entender como se da a formação do valor locativo do imóvel de acordo com um padrão de variáveis, para os dados e a situação estudada.

Na seção 4.1, foi encontrado a presença de dados considerados como Outliers e que foram excluídos de todas as análises. Nas seções 4.2 e 4.3, foram selecionados dois conjuntos de variáveis que se mostraram como importantes para a criação do modelo com melhores parâmetros estatísticos. Para o primeiro modelo, obtido através da análise completa do banco, encontrou-se dez variáveis. Para o segundo modelo, esse feito com a retirada das variáveis dicotômicas, cujo a variável tipologia agrega a informação, foi selecionado um modelo com apenas cinco variáveis.

Na seção referente aos resultados da Árvore de Regressão e Florestas Aleatórias, 4.4, encontrou-se uma relação na formação do valor locativo do imóvel entre a variável Ano de coleta do dado, Público alvo do empreendimento, Área Privativa e População da região do empreendimento e também a relação de importância demonstrando melhores parâmetros para as variáveis Área Privativa, Ano de coleta do dado e Público alvo do empreendimento.

Foi criado e ajustado um modelo com o auxilio do SISDEA que nos permitiu a criação do modelo final com seis variáveis sendo Área Privativa, Estacionamento Privativo, Quantidade de Lojas no empreendimento, Ano de Inauguração do empreendimento, Público alvo do empreendimento e Ano de coleta do dado. O modelo final é composto por variáveis que apresentaram importância nas análises anteriores, como era de se esperar e mostrou-se ser um modelo com características superiores ao demais, através das estatísticas e resultados como por exemplo o coeficiente de determinação que anteriormente estava em torno de 60% e passou para 84% no modelo final, ou seja, o modelo final é capaz de explicar, para o caso R^2, 84% da variável resposta, mais de 20% a mais do que os modelos inicialmente estudados.

Por fim, a junção de técnicas possibilitou a melhor escolha das variáveis, através das análises paramétricas e não paramétricas e a melhor modelagem e definição de um modelo final, através do SISDEA, possibilitando a conclusão de um modelo com graus de precisão e fundamentação 3, dentro das Normas (NBR-14.653-2, 2011), sendo estes os melhores padrões e consequentemente o melhor modelo para a conclusão do valor locativo de um imóvel que se enquadre na comparação com os dados comparativos da amostra.

Referências

ABUNAHMAN, S. A. *Curso básico de engenharia legal e de avaliações*. [S.l.]: Pini, 2008.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BUSSAB, W. d. O.; MORETTIN, P. A. Estatística básica. In: *Estatística básica*. [S.l.: s.n.], 2010. p. xvi–540.

DANTAS, R. A. *Engenharia de Avaliações: uma introdução à metodologia científica*. [S.l.]: Pini, 2005.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.

KREMER, J. *Mercado imobiliário*. [S.l.]: Uniasselvi, 2008.

MATOS, D.; BARTKIW, P. I. N. Introdução ao mercado imobiliário. *Curitiba: Instituto Federal de Educação, Ciência e Tecnologia–Paraná–Educação a distância*, 2013.

NBR-14.653-1. *Avaliação de imóveis urbanos*. [S.l.], 2019.

NBR-14.653-2. *Avaliação de imóveis urbanos*. [S.l.], 2011.

NETO, A. P. Curso de engenharia de avaliação imobiliária– regressão linear e inferência estatística. *Belo Horizonte, MG:[sn]*, 2021.

NETO, A. P. *Pelli Sistemas*. 2022. ⟨https://pellisistemas.com/⟩.

SOUSA, A. A. d. *Atuação do programa de financiamento carta de crédito Caixa no mercado imobiliário. 2006. 172 f.* Tese (Doutorado) — Dissertação (Mestrado em Planejamento Urbano e Regional)–Universidade . . . , 2006.