UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA ESTRUTURAL E CONSTRUÇÃO CIVIL

CURSO DE GRADUAÇÃO EM ENGENHARIA CIVIL

GUSTAVO HENRIQUE PINHEIRO DA SILVA

MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE FORTALEZA

FORTALEZA

2019

GUSTAVO HENRIQUE PINHEIRO DA SILVA

MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE FORTALEZA

Monografia apresentada ao Curso de Graduação em Engenharia Civil do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Engenheiro Civil.

Orientadora: Prof. Dra. Vanessa Ribeiro Campos

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S58m Silva, Gustavo Henrique Pinheiro da.
Modelos de aprendizagem de máquina para precificação de imóveis na cidade de Fortaleza / Gustavo Henrique Pinheiro da Silva. – 2019.
86 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Civil, Fortaleza, 2019.
Orientação: Profa. Dra. Vanessa Ribeiro Campos.

1. Mercado imobiliário. 2. Avaliação de imóveis. 3. Sistemas de Apoio a decisão. 4. Machine-Learning. 5. Estimação de preços. I. Título.

CDD 620

GUSTAVO HENRIQUE PINHEIRO DA SILVA

MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE FORTALEZA

Monografia apresentada ao Curso de Graduação em Engenharia Civil do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Engenheiro Civil.

Aprovada em: 25 de novembro de 2019

BANCA EXAMINADORA

Prof. Dra. Vanessa Ribeiro Campos (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Anselmo Ramalho Pitombeira Neto
Universidade Federal do Ceará (UFC)

Eng. Sandro Ricardo Vasconcelos Bandeira
Secretaria Municipal de Finanças - Fortaleza (SEFIN)

À minha mãe, Eleneida; a minha irmã, Sheila; a meus sobrinhos, Guilherme Lucas, João Gabriel e João Pedro.

AGRADECIMENTOS

Á Universidade Federal do Ceará que foi minha casa nesses últimos 5 anos de graduação e me proporcionou conhecer algumas das mentes mais brilhantes com as quais tive oportunidade de trabalhar.

A minha Orientadora Prof. Dra. Vanessa Ribeiro Campos que acreditou na minha capacidade e aceitou me orientar sobre tema tão complicado e importante. Ao professor Prof. Ms. José Ademar Gondim Vasconcelos por despertar meu interesse pelo mercado imobiliário. A Prof. Dra. Marisete Dantas de Aquino por se disponibilizar para ler esse projeto e prestigiar a apresentação final deste fechamento de curso.

A Secretaria de Finanças do Município (SEFIN), pela disponibilização da base de dados, sem a qual este trabalho não poderia ter sido realizado.

Ao Prof. Dr Anselmo Ramalho Pitombeira Neto junto com o restante da minha banca de avaliadores, pelas valiosas contribuições, considerações e experiências trazidas a este trabalho

Ao Centro de Empreendedorismo da UFC que não só acendeu a grande fagulha do pensar fora da caixa, como mostrou o quanto era possível fazer quando realmente desejamos algo e por ter me permitido criar um projeto enorme enquanto gestor.

Ao Prof. Dr. Abraão Freire Saraiva Júnior do departamento de engenharia de produção, amigo que não só deu vários puxões de orelha como também incentivou e ensinou muito do que aprendi nesses últimos anos.

Ao Prof. Dr. Tibérius de Oliveira e Bonates do Dep. de Estatística e Matemática Aplicada e ao mestrando em engenharia Civil Felipe Fernandes Moreira, por me darem suporte e ajudarem a melhorar as minhas premissas para estabelecer meus modelos finais.

A minha mãe por ter me dado o suporte e o tempo que eu precisava para me dedicar não só a esse trabalho, mas durante todo o tempo da graduação, por acreditar em mim e por me incentivar a ser o primeiro filho formado. Ao meu pai por ter me ajudado financeiramente por todo esse tempo de vida, e a minha irmã e sobrinhos que sempre foram um ponto fora da curva na minha rotina.

Um agradecimento especial a minha atual Companheira e Futura Engenheira Civil Brenda Arielly Mendonça Rodrigues por estar sempre comigo quando as coisas pareciam que não iam dar certo e por todo o carinho lendo e relendo este trabalho para que ele pudesse sair da melhor forma possível.

A todos os professores que passaram pela minha vida, sejam eles da Universidade Federal do Ceará ou do Instituto Federal do Ceará, por terem deixado suas contribuições para que eu pudesse chegar ao final desse trabalho.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

A você que está lendo este trabalho, pessoa que dá sentido a esta escrita e que espero que consiga usá-lo para impulsionar nosso mercado imobiliário.

“Acredite e não se explique pois poucos vão entender: só se compreende um sonho se o sonhador for você.“

(Bráulio Bessa)

RESUMO

As transações imobiliárias são realizadas com a participação de diversos atores. Muito para além do comprador e vendedor, financiadores e órgãos governamentais atuam e são diretamente impactados com seu resultado e efetividade. Por representar quase sempre um valor da ordem de dezenas ou centenas de milhares de reais, todos os atores sempre visam a minimização do risco que sofrem nessa modalidade de operação. Muito antes dos estudos de concessão de crédito, os estudos sobre precificação do imóvel podem levar ou não à alienação do bem. As metodologias de avaliação previstas em normas técnicas abordam principalmente aspectos de vistoria ou, em alguns casos a análise estatística do imóvel e sua circunvizinhança. Com o avanço da computação e das teorias estatísticas, surgiram novos métodos de avaliação, como a aprendizagem de máquina que vem demonstrando bons resultados como oráculo de valores atingindo, em alguns casos, a erros médios próximos ou menores a 6%. Fatores de erro tão baixos tornam-se bastante impactantes no mercado, pois trazem segurança para todas as partes durante o processo aquisitivo. Este trabalho propõe a aplicação das metodologias de aprendizagem computacional assistida para avaliação de imóveis, discutir seus resultados e escolher qual delas apresenta melhores resultados para a cidade de Fortaleza. Os dados para a pesquisa foram fornecidos pela Secretaria de Finanças do Município e contêm informações sobre valores das transações e dados físicos da propriedade e serão o ponto de partida para cada um dos modelos de treinamento e previsão abordados no corpo deste trabalho.

Palavras-chave: Mercado imobiliário. Avaliação de imóveis. Sistemas de Apoio a decisão.*Machine-Learning* . Estimação de preços.

ABSTRACT

Real estate transactions are carried out with the participation of various agents. Far beyond buyers and sellers, funders and government agencies impact and are impacted by their result and effectiveness of those transactions. All agents always aim to minimize the risk they face in this mode of operation since it represents a sum of about tens or hundreds of thousands of reais. Long before credit studies, real estate pricing studies may or may not lead to the sale of the property. Evaluation methodologies present in technical standards mainly address inspection aspects or, in some cases, the statistical analysis of the property and its surroundings. With the advances in computer science and statistical theories, new evaluation methods have emerged, such as machine learning, which has shown good results as oracle of values, reaching, in some cases, average errors close to or less than 6%. Such low error rates become quite influential in the market since they bring security to all parties during the purchasing process. This paper proposes to discuss the insertion of several assisted computational learning methodologies to real estate valuation, discuss their results and choose which one presents the best results for the city of Fortaleza. The data for the survey were provided by the Municipal Finance Department of Fortaleza. They contain information on transaction values and the real estate physical data, which will be the starting point for each of the training and forecasting models addressed in the body of this paper.

Palavras-chave: Real State Market,Real estate assessment. Decision support systems. Machine-Learning. Price Valuation

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

Bagging	Bootstrap Aggregating
GB	*Gradient Boosting*
GBM	*Gradient Boosting Machine*
GLM	*General Linear Model*
IPTU	Imposto Predial e Territorial Urbano
IRLSM	*Iteratively Reweighted Least Squares Method*
ITBI	Imposto de Transmissão de Bens Imobiliários
LUOS	Lei de Uso e Ocupação do Solo
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentual Error*
MLP	*Multlayer Perceptron*
MSE	*Mean Squared Error*
nfolds	*Number of Folds*
RF	*Random Forests*
RLM	*Regressão Linear Múltipla*
RMSE	*Root Mean Squared Error*
RMSLE	*Root Mean Squared Logarithmic Error*
RNA	Redes Neurais Artificiais
SEFIN	Secretaria Municipal de Finanças do município de Fortaleza

SUMÁRIO

1 INTRODUÇÃO

A avaliação de imóveis é uma prática normatizada pela NBR 14653 (ABNT, 2011). Tratado como bem, a prática nesse tipo de empreendimento é definida como uma análise técnica, realizada por um engenheiro de avaliações, para identificar o valor de um imóvel, de seus custos e de frutos diretos.

Este trabalho tem como objetivo propor, por meio de uma análise do mercado imobiliário da cidade de Fortaleza-CE, um modelo de previsão para a precificação de imóveis residenciais. A metodologia será definida a partir de dados de transações imobiliárias da SEFIN, consistindo na aplicação de 5 modelos de aprendizagem diferentes, sendo eles: Regressão linear, Regressão Gaussiana, *Random Forests* , *Gradient Boosting Machine*, Redes Neurais Artificiais, este último já descrito na NBR 14653-2 a partir do ano de 2011 e o primeiro sendo o modelo mais praticado atualmente no mercado. A comparação com outros modelos ainda não descritos em norma vem como forma de instigar o uso de novas metodologias estatísticas a fim de fornecer um modelo mais adequado de predição dos valores a serem praticados pelo mercado. Os resultados serão confrontados com os outros para a verificação daquele que possui melhor acurácia. O processo de desenvolvimento do trabalho ocorreu inicialmente através do processamento de dados, da escolha das variáveis significativas e, em seguida, do uso das metodologias *machine-learning* e regressão.

Para a verificação da eficácia, os dados serão validados com o modelo *k-Fold* de análise cruzada, com algumas análises comparativas de comportamentos de modelos, como:*Mean Absolute Error* (MAE),*Mean Squared Error* (MSE) ,*Root Mean Squared Error* (RMSE), *Root Mean Squared Logarithmic Error* (RMSLE) e, por fim, o *Mean Absolute Percentual Error* (MAPE), obedecendo ao nível máximo, descrito em norma, um nível de erro igual 15%. Tornando, desse modo, possível uma análise acurada para os mais diversos atores, como compradores, vendedores e financiadores da aquisição. Neste trabalho, os dados utilizados foram referentes ao fechamento dos contratos de unidades unifamiliares residenciais.

1.1 Justificativa

Vários dilemas sempre se puseram na avaliação de imóveis, sendo os principais problemas com o comércio desse bem durável causados pela baixa liquidez decorrente de uma demanda extremamente volátil de compra. A análise de imóveis se dá por meio de uma

precificação hedônica, que leva em consideração tanto fatores internos quanto externos ao bem avaliado e muitas dessas variáveis tem o comportamento qualitativo, por isso tornam-se difíceis de terem seu valor mensurado. Para a precificação das variáveis qualitativas, usam-se metodologias estatísticas de comparação de preços de imóveis, sendo os mais comuns os modelos de regressão linear. No modelo linear (ou no modelo de regressão múltipla) a variável dependente é expressa como combinação linear das variáveis independentes acrescidas de um erro aleatório. Embora simples, os modelos precisam, por definição, que os dados mantenham o comportamento aproximadamente igual a uma função de primeiro grau, o que nem sempre é notado na prática e pode resultar em previsões menos representativas.

Todo empreendimento imobiliário representa uma soma considerável de dinheiro e, portanto, é de interesse de todos os atores que se proponha um valor adequado de venda: melhorando o lucro do vendedor, diminuindo o risco do financiador e gerando um preço justo ao cliente. Um valor superdimensionado tanto poderá diminuir a procura pelo bem, quanto, mesmo com o interesse do comprador e vendedor, dificultar a obtenção de financiamento da transação.

A aprendizagem assistida surge como a melhor discriminante para decidir entre duas classes com distribuições normais e equivalentes matrizes de covariância (BISHOP, 1995). A técnica de Redes Neurais Artificiais surge como um modelo baseado na estrutura biológica humana de funcionamento neuronal e é vantajosa ao analisar grandes quantidades de dados para treinar e fazer uma previsão do modelo. O modelo ganha destaque por dispensar a necessidade da linearidade das variáveis, devendo-se ter cuidado apenas para controlar a variância dos resultados, dadas variáveis de calibração. Além disso, devido sua capacidade adaptativa, é capaz de responder a falhas de premissas ou de tratamentos, podendo ser usado para sintetizar uma grande superfície de decisão com o mínimo de pré-processamento (Lecun *et al.*, 1998). Aliados a esse, embora menos abordados, temos outros preditores singulares ou mesmo modelos conjuntos de inteligência artificial, cada um com suas vantagens, tentando gerar um valor mais próximo à realidade. Com um preço mais assertivo, mais propriedades podem ser negociadas, visto que os agentes financeiros possuem mais confianças para ofertar melhores condições de pagamento, aumentando assim a movimentação de dinheiro no setor como um todo.

Atualmente, as metodologias de aprendizagem de máquina têm ganhado destaque nas mais diversas áreas: análise de comportamentos e padrões; a produção de conteúdo midiático, reconhecimento de padrões, robótica e modelos de previsão econômica. Embora distintas em metodologias, a grande parte das previsões via *machine-learning* possui a vantagem de trabalhar

com modelos mais complexos e que não apresentam comportamento bem definido linearmente. Embora atualmente o uso de redes neurais seja o único diretamente previsto na (ABNT, 2011), o modelo traz algumas desvantagens como alta variância de resultados, dado este que inclusive é citado na própria normatização. Outros modelos, por sua vez, apresentam menor variabilidade, o que pode resultar no final em um modelo que possa trazer melhores resultados.

Mesmo com metodologias já bem definidas, a área da aprendizagem assistida continua inovando, um de seus principais usos atuais é o dos algoritmos *ensemble* que surgiram como forma de balancear a "votação"que levará ao resultado final apresentado com o uso de múltiplos *learner's* distintos, que podem acabar se compensando em suas falhas.

Entretanto o modelo do mercado vem mudando constantemente, fatores externos à intenção de compra e venda do bem, como econômicas, políticas ou normativas ocasionam mudanças bruscas no mercado da construção. Conclui-se então que mesmo chegando a um modelo de melhor previsibilidade, é necessário que se dê continuidade aos estudos, alimentando o modelo com os novos dados ou alterando-o para algum que se adéque melhor a realidade do local ou do tempo analisados.

Nesse estudo, não será abordado o campo da arbitrariedade, trabalhará apenas com dados reais dos imóveis. Os dois tipos de variáveis (quantitativas e qualitativas) serão levados em consideração quando possível. Para modelos mais simples como regressão linear, variáveis categóricas não serão usadas.

1.2 Objetivos

Esse trabalho tem por objetivo propor o melhor modelo dentro os analisados que seja capaz de calcular o preço das unidades habitacionais unifamiliares da cidade de Fortaleza com a aplicação de aprendizagem de máquina. Entre os objetivos específicos têm-se:

- Verificação da representatividade e escolha das variáveis usadas
- Treinamento, ajuste e calibração com algoritmo de regressão linear múltipla
- Treinamento, ajuste e calibração com algoritmo de regressão Gaussiana
- Treinamento, ajuste e calibração com algoritmo RF
- Treinamento, ajuste e calibração com algoritmo GBM
- Treinamento, ajuste e calibração com algoritmo RNA
- Verificação da precisão do resultado com o Uso de algoritmo *k-fold* e erro médio percentual
- Escolha final do modelo de criação das previsões

2 REVISÃO BIBLIOGRÁFICA

2.1 O Mercado Imobiliário

2.1.1 Formação de preços

O mercado imobiliário é considerado totalmente diferente dos outros, tanto para imóveis quanto para terrenos, porque cada produto é único e de alta duração (DIPASQUALE; WHEATON, 1996). No mercado de celulares, automóveis ou outros produtos manufaturados temos partes do todo parecidas, o que possibilita a substituição de componentes parecidos por outros que se comportarão da mesma forma. O fator independência entre transações torna difícil a avaliação, dadas as determinadas características de um imóvel. Entretanto, é sabido que algumas variáveis podem ser usadas para a precificação, a exemplo nota-se o local da edificação: os imóveis mais próximos de centros comerciais ou com uma boa infraestrutura pública têm uma valoração externa, outros fatores como área total do terreno ou tamanho da testada são importantes na hora de decidir pela aquisição de um imóvel e portanto, um aumento no preço.

Para a avaliação de imóveis, são levadas em consideração, segundo ABNT (2011):

1. Aspectos gerais da edificação, como condições econômicas, políticas e sociais, além dos usos anteriores atípicos e estigmas;
2. Aspectos físicos, como condições de relevo, natureza do terreno e condições ambientais;
3. Localização;
4. Uso e ocupação do solo, como descrito na Lei de Uso e Ocupação do Solo (LUOS);
5. Infraestrutura do local, como sistemas de transportes, saneamento básico, energia elétrica, sistemas de comunicação e outros itens de comunidade;
6. Aspectos construtivos, arquitetônicos, paisagísticos e culturais, benfeitorias ou a presença de patologias ou avarias na edificação;

2.1.2 O Mercado Brasileiro

Em 2008 ocorreu a bolha especulativa no mercado imobiliário dos Estados Unidos, razão que levou ao aumento do interesse de especialistas por analisar o mercado imobiliário de diversos países. O governo brasileiro, em ação para mitigar os efeitos da crise, reduziu as regras de concessão de crédito para aumentar o financiamento nos mais diversos setores. No mercado imobiliário, a flexibilização das regras de crédito causou um fenômeno que aumentou

expressivamente o preço dos imóveis. Além disso o período estável na economia, que veio a seguir com estabilização de preços, a queda na taxa de juros e a expansão das obras públicas mirando a Copa do Mundo de Futebol de 2014, trouxe confiança para aumentar o número de aquisições desses bens. (ELDIONARA *et al.*, 2014).

Após o processo de valoração, o país entrou em clima de recessão com aumento do desemprego e diminuição da renda, impactando negativamente na prévia análise de crédito feita nos anos de altos investimentos no mercado de capital (RIBEIRO; BERTRAN, 2019). A alta oferta de crédito imobiliário criou um mecanismo que transferiu o dinheiro da poupança para incorporadoras e bancos como entrada na compra de construções. Entretanto com o advento da crise, tornou-se impossível, para muitos casos, manter o parcelamento acordado, fator esse determinante para que no período houvesse o aumento da alienação fiduciária e da quebra de contratos com as incorporadoras. Estimam-se que, aproximadamente, a cada dois imóveis vendidos pelas maiores incorporadoras no Brasil, um foi "devolvido"entre os anos de 2015 (PASQUALIN, 2016), causando uma perda na pretensão aquisitiva da população, além da perda permanente dos valores investidos pelos consumidores, visto que apenas uma parte do montante até então investido era recuperada pelos contratantes. A importância da boa medição dos preços torna-se imprescindível para garantir uma boa negociação do imóvel e uma boa avaliação de crédito para uma transação.

2.1.3 Métodos tradicionais de avaliação

Para a ABNT, quando se deseja aferir o valor de mercado de um imóvel, deve-se, sempre que possível, preferir o método comparativo direto com o mercado, sendo recomendável também a apresentação das considerações quanto ao aproveitamento eficiente do imóvel (ABNT, 2011).

Inicialmente deve-se planejar a pesquisa mercadológica da região, esta deve visar obter dados de mercado com características semelhantes quanto ao que se pretende avaliar. A pesquisa deve caracterizar e delimitar o mercado de análise e nele as variáveis dependentes e independentes, que serão usadas nas etapas seguintes para estabelecer o valor do bem.

Preliminarmente, é saudável que dentro da pesquisa os dados sejam sumarizados e apresentados em forma de gráficos de distribuições de frequência juntamente com as relações entre variáveis para que haja a análise da interdependência dos pontos de pesquisa.

Dadas as vaiáveis, temos duas opções tradicionais de tratamento: por fatores (ho-

mogeneização por fatores e critérios) ou científico (levando em consideração o tratamento das evidências empíricas com metodologia cientifica). Entretanto, temos a exceção do campo do arbítrio que pode ser usado para explicar variáveis que são importantes, mas não puderam ser determinadas estatisticamente. O máximo valor arbitrado tem uma tolerância de 15%.

Para a composição do estudo econômico, podemos seguir pelo método involutivo que leva em consideração:

1. Vistoria;
2. Projeto hipotético;
3. Pesquisa de valores;
4. Previsão de receitas;
5. Levantamento do custo de produção e do projeto hipotético;
6. Previsão de despesas adicionais;
7. Margem de lucro do incorporador;
8. Prazos;
9. Taxas.

Método de Vendas:

1. Estimação das receitas e despesas;
2. Montagem do fluxo de caixa;
3. Estabelecimento da taxa mínima de atratividade;
4. Estimação do valor do imóvel.

Método Evolutivo: a composição de valores é obtida pela conjugação de métodos a partir do valor do terreno e consideradas as benfeitorias e o fator de comercialização.

$$VI = FC(VT + CB) \tag{2.1}$$

Onde VI representa o valor do imóvel, VT o valor do terreno, CB o custo da reedição da benfeitoria e FC o custo da comercialização.

Quanto aos comparativos direto de dados do mercado a norma recomenda os modelos de múltipla regressão linear especificando três graus de fundamentação, variando de I a III, onde III representa o laudo mais completo e ideal para a tomada de preços.

Entre outros modelos tomados pela NBR14653-2, começamos a notar o surgimento das metodologias de aprendizagem de máquina. Sendo capazes de prever não apenas linearmente, mas assistidamente usar modelos recursivos para efetuar uma predição mais próxima da realidade.

2.2 Métodos de aprendizagem

A análise estatística é a principal ferramenta para comparação do imóvel diretamente com outros do mercado. Algumas abordagens mais consagradas tentam descrever essa relação com modelos lineares de aproximação, outras mais recentes utilizam do aprendizado assistido como forma para discretizar determinado mercado.

As metodologias de aprendizagem de máquina são definidas através de regras matemáticas distintas, têm efetividade pois tentam aliar a capacidade de processamento computacional com regras de decisão humanas. A possibilidade de analisar variáveis independente de correlação direta e a precisão dada para modelos de previsão complexos são destaques que as fazem ser usadas nos mais diversos segmentos que vão de análises financeiras até processamento de imagens. As metodologias diferentes se tornam as avaliações mais adequadas para análises qualitativas como no caso das *Random Forests* e GBM, ou as quantitativas como as Redes Neurais artificiais e metodologias como GLM,*boosting*. Com a evolução da bibliografia, chegamos aos modelos de classificação conjuntos, responsáveis por agrupar preditores fracos com os preditores mais fortes.

2.2.1 Regressão Linear

O modelo mais usado quando se desejar calcular o comportamento de uma variável dependente é o de regressão linear, onde através de combinação linear, unimos as variáveis independentes do modelo com a finalidade de gerar as variáveis dependentes (ABNT, 2011).

$$P = a_1 x_1 + a_2 x_2 + ... + a_n x_n \tag{2.2}$$

Para se usar esse tipo de modelo, é necessário garantir que a amostra apresente alguns princípios básicos como normalidade, homoscedasticidade, não-multicolinearidade, não-autocorrelação, independência e inexistência de pontos atípicos. Quando as observações apresentam todas as características necessárias, podemos propor um modelo de regressão. Outro fator importante é que o imóvel previsto necessariamente tem que ter características coerentes com a multilinearidade referida.

O modelo, embora muito usual, falha quando a amostra apresenta um comportamento não normal ou um baixo número de observações, sendo necessário o uso de métodos com funções genéricas de avaliação.

2.2.2 Meta Algorítmo Bagging

A técnica de Bootstrap Aggregating (Bagging) propõe uma substituição na lista de dados usada inicialmente para o treinamento da rede. Ao invés de usarmos um modelo de predição $\phi(x,£)$, sendo x a variável preditora e $£$ o modelo treinado, pode ser substituído com melhor acurácia por um modelo $\phi(x,£k)$ onde $£k$ seria representado pela média das subamostras $£$. Em suma, propor um novo modelo $\phi a(x) = E£\phi(x,£)$, onde E denota o valor esperado e $\phi a(x)$ denota agregação (LEO, 1996).

Usualmente replica-se a amostra $£$ em k subamostras, denominadas *bootstraps*, retiradas aleatoriamente com reposição, onde cada subamostra possui uma quantidade de elementos expressiva (70% até 100%) em relação a amostra principal, garantindo a repetição de elementos retirados em cada *bootstrap*. Quanto ao modelo, após simulados k *bootstraps*, agregam-se cada um desses modelos com igual peso, fazendo-se uma ponderação sobre o valor final previsto. O modelo por si consegue reduzir em média 30% do erro final.

2.2.3 Meta Algorítmo Boosting

Ao contrário da técnica de *bagging*, a técnica de *Boosting* não usa o modelo aleatório para gerar os *bootstraps*. Este modelo baseia-se no princípio da aprendizagem fraca, cada nova amostra deve ser, pelo menos, um pouco melhor que a outra, um modelo que gera interações que vão se aperfeiçoando, não sendo obrigatório que ele seja bom na primeira tentativa (SCHAPIRE, 1990).

A primeira amostra é tomada aleatoriamente sem restrições e usada para o primeiro modelo de previsão probabilístico, após a primeira simulação compara-se o modelo previsto com o real amostrado procurando os casos em que o modelo de predição falha. As observações falhas são colocadas dentro do novo *bootstrap* e gera-se a observação N + 1, novamente a amostra é simulada e é calculado o novo erro. O processo interativo do *Boosting* repete até N + 1 = k, onde k é o número designado de interações.

Outra diferença para método de *bagging* é que o modelo final de previsão não representa somente o valor mais votado ou a média dos valores das simulações, o valor predito final representa uma média ponderada entre as simulações de cada *bootstrap* pela acurácia individual de cada um dos modelos

$$\phi a(x) = Pr(x)\phi(x,£) \tag{2.3}$$

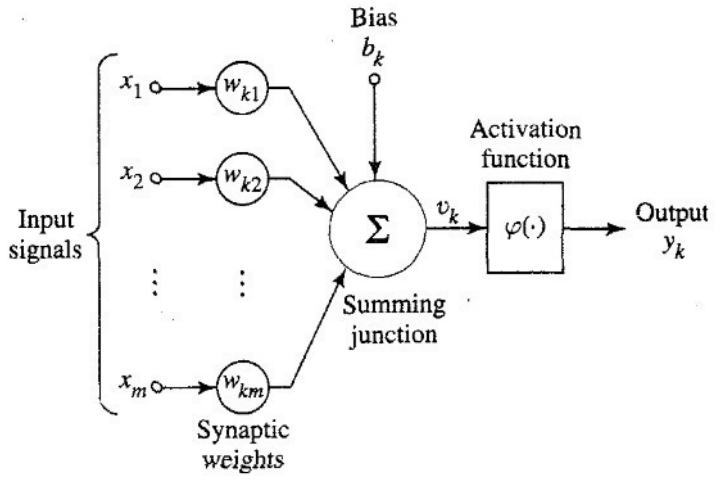
Para Pr(x) dado como a probabilidade de se ter um classificador difícil. O modelo foi reconhecido pela sua capacidade de adicionar modelos fracos de aprendizagem para gerar um modelo de previsão melhor.

2.2.4 Redes Neurais Artificiais

As Redes Neurais Artificiais surgiram como uma alternativa de modelo preditivo baseado no comportamento humano. O cérebro humano é tido como a estrutura computacional mais completa até então descoberta pelo homem, é uma estrutura não linear, complexa e um sistema de informação paralelo de processamento de dados. Composta por neurônios ela é ainda a estrutura mais veloz para reconhecimento de padrões variados (HAYKIN, 1994).Sua principal vantagem é a de efetuar a previsão de modelos não lineares de correlação. Entretanto, a mesma vantagem traz a possibilidade de haver uma maior dispersão nos resultados, dependendo principalmente da base usada para o treinamento do modelo, devendo-se respeitar apenas um comportamento próximo a normalidade. Por conta da variância a (ABNT, 2011) recomenda o uso de algoritmos de treinamento amostral como *Bagging* ou multiobjetivo.

As Redes Neurais imitam o comportamento cerebral pelas suas divisões em neurônios, denominados perceptrons, sendo divididas entre o uso de um mais perceptrons (divisão dependendo do número de variáveis abordadas), dispostos na seguinte ordem: Nós de entrada , conexões ponderadas de nós aos perceptrons (equivalentemente aos neurônios e sinapses do cérebro humano) e uma função de ativação. como disposto na Figura 1.

Figura 1 – Estrutura de uma rede neural



Fonte: Haykin (1994)

As redes mais comuns são as *Multlayer Perceptron* (MLP), que na etapa dos nós de entrada possui uma série de sensores, representando o comportamento agregado de múltiplas variáveis no modelo preditivo, a análise de imóveis é praticamente atrelada a esse modelo devido à complexidade do modelo de causalidade, sendo os sensores mais comuns: Número de quartos, área total, área construída, número de banheiros e quantidade de vagas para carros, entre outros. Após imputados os valores, passamos pelas sinapses, que são um *layer* oculto contemplando os pesos recalibráveis para a função de ativação da saída (HAYKIN, 1994), A rede final pode ser calculada fazendo a soma dos pesos de cada um desses sensores nos primeiros nós; após passado o modelo pela camada oculta, os valores são usados como entrada para uma função sigmoide denominada função de ativação que normalizará os valores fornecendo o resultado final. O peso conectando a entrada i até o neurônio oculto j é chamada de W_{ij}, já o peso conectando o neurônio oculto j para o neurônio de saída k é chamado de V_{kj}.

Para os neurônios ocultos:

$$net_j^h = \sum_{i=1}^{N} W_{ji} x_i e y_i = f(net_j^h) \tag{2.4}$$

Pra os neurônios de saída:

$$net_j^o = \sum_{j=1}^{J+1} V_{kj} y_i = f(net_k^0) \tag{2.5}$$

A principal vantagem das redes neurais é a de ser um modelo recursivo que trabalha em cima da camada oculta reponderando os pesos, alterando o valor levado para a função de ativação, o que por sua vez altera o valor final. Para fazer essa alteração, deve-se calcular o erro final que o modelo traz, comparando o valor previsto com os valores colocados para treinamento. O objetivo final da etapa de recalibração é o de reduzir o erro geral, para tanto usa-se a função:

$$W_{\mathrm{ij}}(t+1) = W_{\mathrm{ij}}(t) + c\lambda^2 y_j(1-y_j)x_1(t)(\sum_{k=1}^{k}(d_k - o_k)o_k(1-o_k)v_{kj}) \tag{2.6}$$

O modelo prosseguirá até o momento que seja atingida a precisão desejada ou o número de repetições fornecidas. Dependendo da complexidade do modelo ou do número de repetições adotadas, algumas simulações podem demorar dias, semanas ou meses. Um revés do modelo é a alta variância da saída, o que o torna ainda de difícil aceitação para avaliação no mercado imobiliário. Para diminuir a variância das RNA's, poderíamos usar calibragem por uma pré amostragem, usando o método de *bagging* disposto na seção 2.2.2 ou trabalhar com um grande número de dados.

2.2.5 General Linear Model (Gaussian Regression)

O modelo linear geral de regressão Gaussiana, é um modelo linear onde aproximamos nosso modelo de previsão por uma função de gauss. Ela é definida por uma Gaussiana com N observações e p preditores, com cada par de preditores tendo a mesma correlação populacional ρ, com ρ variando de 0 a 0,95. Os valores são gerados pela função.

$$Y = \sum_{j=1}^{\rho} X_j \beta_j + k * Z \tag{2.7}$$

Com $\beta_j = (-1)^j exp(-2(j-1)/20), Z \sim N(0,1)$, e k escolhido mantendo a relação sinal-ruído = 3. Os coeficientes foram construídos de forma a obter sinais alternantes e valores exponencialmente decrescentes (FRIEDMAN *et al.*, 2010).

2.3 Ensembles Classifiers

A combinação de classificadores ou *Ensembles Classifiers* consiste na tentativa de se obter melhores predições através de classificadores mais fracos. Surgiram basicamente como modelo para melhorar a acurácia das aproximações individuais de previsão (TSAI *et al.*, 2014).

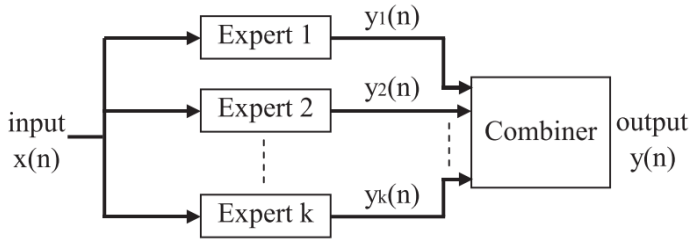
Esse tipo de abordagem apresenta menor eficácia para bons preditores, não deixando, no entanto, de ser relevante para diversas aplicações como bancos ou outras operações financeiras, visto que previsões mais acuradas podem impactar significativamente nos lucros dessas organizações. Esses classificadores podem ser dispostos em dois tipos: o *parallel ensemble* e o *sequential ensemble*, sendo inclusive denominada como uma metodologia avançada de previsão imobiliária segundo Selim (2009).

2.3.1 Parallel Ensemble

O método modular é inspirado no cérebro humano que pode se dividir em tarefas distintas sem interferência mútua nos resultados, cada um resolvido por um oráculo. A razão desse tipo de abordagem se dá principalmente devido à redução do erro total quando se estabelece uma avaliação final em cima da média de cada um dos avaliadores parciais. A partir da divisão paralela temos um banco de dados novo, remontado a partir das análises individuais de cada subtarefa. A combinação paralela é usada principalmente usado para classificadores fracos. A

Figura 2 representa o modelo gráfico que produz o *output* final do modelo.

Figura 2 – Modelo de um classificador *parallel ensemble*



Fonte: Tsai *et al.* (2014).

2.3.2 Sequential Ensemble

Sua principal função é explorar a dependência entre os meta-algoritmos dos métodos de aprendizagem, a precisão pode ser melhorada recalibrando o resultado do modelo de previsão fraco com a utilização de um modelo de previsão forte, agrupando os dados, seu funcionamento se dá na forma:

1. Especificar a lista de algoritmos base;
2. Especificar o algoritmo de meta-aprendizado.

Treinamento do algoritmo ensemble:

1. Treinamos os L algoritmos base;
2. Usa-se o algoritmo *k-Fold* de validação e coleta-se os valores para cada um dos L algoritmos;
3. Forma-se uma matriz contendo com o número de linhas de determinada base de treinamento e os preditores aferidos.

Predição de novos dados:

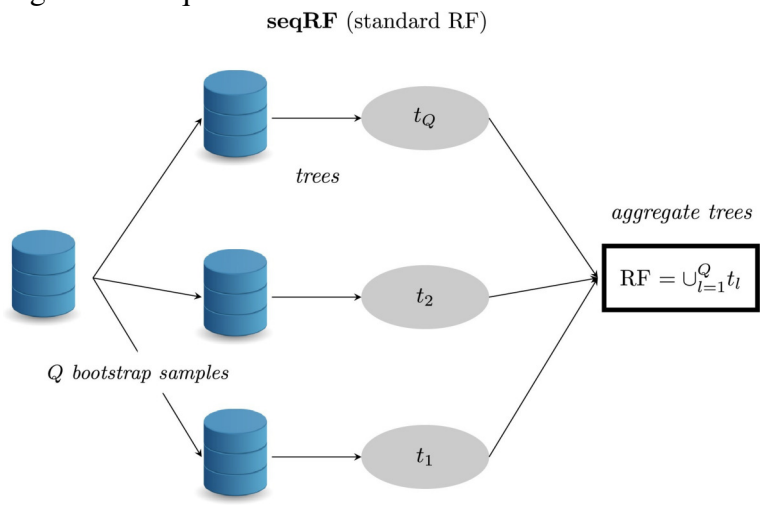
1. Gerar algoritmo que trabalhe em cima das bases de aprendizado;
2. Juntar todas as funções na função de meta aprendizado final.

2.3.2.1 Random Forests

O modelo de florestas aleatórias foi abordado inicialmente como generalização dos modelos de árvore de decisão. Consiste em um classificador $[h(x,\theta_k), k = 1....]$ onde θ_k representa um conjunto de vetores aleatórios tomados independentemente e X uma sub-amostra, saída de um algoritmo de *bagging* ou apenas uma sub-amostra do modelo inicial. Ademais, cada

arvore poderá votar uma vês no valor que mais lhe couber para a entrada X (BREIMAN, 2001).

Figura 3 – Esquemático de uma *Random Forest*



Fonte: Genuer *et al.* (2017)

.

O valor final de saída do modelo será o mais representativo votado pelas n árvores da simulação. Ele possui vantagem sobre as árvores de decisão definidas pela premissa de que o ser humano é incapaz de escolher a melhor arvore de decisão para o modelo, fator este mitigado pela grande quantidade de árvores definidas aleatoriamente.

O modelo pode ser usado tanto para classificação quanto para regressão e trabalha bem com variáveis classificatórias e contínuas, sendo o primeiro caso o seu modo mais utilizado.

2.3.3 GBM

O *GBM* é um modelo *Sequential Ensemble* que se baseia nos critérios de *Boosting* e RF para avaliar as amostras. É executado um modelo de *Boosting* como entrada para uma floresta aleatória, em seguida é estabelecida uma função gradiente para ajudar na convergência do modelo ao resultado. Ele é otimizado porque, ao invés de estabelecer pesos para cada classificador fraco, ela atribui uma função que incide diretamente sobre a "função perda"do preditor, tentando minimizar a diferença entre o real e o previsto (FRIEDMAN, 2001). Ou seja:

$$F^* = argminE_{y,x}L(y,F(x)) \tag{2.8}$$

onde $L(y,F)$ representa a função perda que deve ser minimizada.

2.4 Validação de Resultados

2.4.1 Método K-Fold

Os modelos de previsão usam o aprendizado do modelo para estimar valores reais, entretanto essa estimativa costumeiramente produz divergências do valor esperado. Por conta desse comportamento, estimar erros é essencial para verificar a aderência de um modelo.

O método *k-Fold* de validação cruzada parte do banco de dados de testes D. Os dados são inicialmente divididos em subconjuntos mutualmente exclusivos, D1, D2, D3 .... D_n com aproximadamente o mesmo tamanho. Em seguida, o oráculo é treinado k vezes (motivo do nome *k-Fold*). O validador cruzado estima a acurácia a partir do número de classificações corretas dos subconjuntos amostrais tomados em relação ao total de classificações dos D subconjuntos (KOHAVI, 1995).

$$acc_{cv} = \frac{1}{n} \sum_{v_i, y_i \varepsilon D} \delta(I(\frac{D}{D_{(i),} v_i}), y_i) \tag{2.9}$$

Cada subconjunto D_i é tomado aleatoriamente e ao fim temos um avaliador estatístico do modelo testado e, repetindo as simulações com um número diferente de *folds*, geramos uma melhor simulação de Monte-Carlo.

A metodologia *k-Fold* retorna as validações em termos de algumas medidas, suas definições matemáticas foram retiradas de seus algoritmos preditores: MSE, RMSE, MAE, RMSLE preditos no módulo de performance do H2O (H2O.AI, 2019) e MAPE no módulo MLMetrics (CHEN, 2017)

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.10}$$

$$RMSE := \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{2.11}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y| \tag{2.12}$$

$$RMSLE = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \ln(\frac{y_i + 1}{\hat{y}_i + 1})^2} \tag{2.13}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|(y_i - \hat{y}_i|}{y_i} \tag{2.14}$$

Para:

N = Número de observações

y = Valor Real

$\hat{y}$ = Valor Previsto

2.5 Aplicações

Os modelos de alienação imobiliária por modelos assistidos também já foram alvo de outros estudos no país, alguns focados em um modelo específico, outros voltados para outros tipos de imóveis, como apartamentos ou terrenos. Apresentaremos alguns desses trabalhos com aplicações diretas da aprendizagem de máquina comoo oráculo preditor.

2.5.1 Predição de preços de imóveis através de aprendizagem de máquina

O trabalho proposto por Malere *et al.* (2019) mostra um modelo de aprendizagem de máquina voltado para todas as regiões do país. Foi usada uma técnica de mineração de dados para a obtenção dos *imputs*. O trabalho também comparou múltiplos modelos, e ao final escolheu o *Gradient Boosting* (GB) implementado pelo pacote "XGBoost"como o melhor modelo.

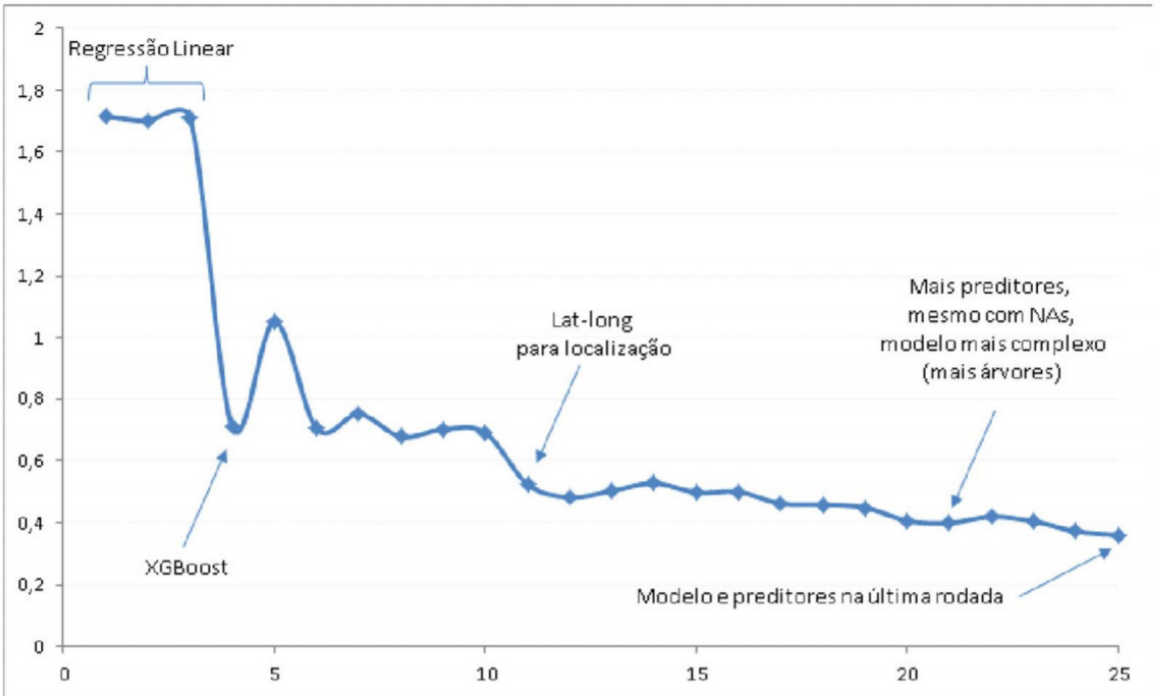
Como variáveis de entrada foram usadas: *Id*, *property_id*, *created_on*, *operation*, *property_type*, *place_name*, *place_with_parent_names*, *country_name*, *state_name*, *geonames_id*, *lat_lon*, *lat*, *lon*, *currency*, *surface_total_in_m2*, *surface_covered_in_m2*, *floor*, *rooms*, *expenses*, *description*, *title*, *image_thumbnail*, *collected_on*, *price*

Após normalizar os dados, efetuar a remoção de *outliers* e submetê-los aos modelos, foi feita a medição de acurácia usando o indicador RMSLE. A Figura 4 mostra a evolução do modelo ao longo de 25 rodadas de simulação, obtendo-se um RMSLE de aproximadamente 0,38.

.

O modelo foi tratado estatisticamente, o que proporciona melhores resultados de previsão. Entretanto, devido a única métrica adotada não ser normatizada não é possível averiguar se o modelo se torna ou não adequado para avaliar o mercado.

Figura 4 – Resultados obtidos nas submissões para trabalho relacionado A



Fonte: (MALERE *et al.*, 2019)

2.5.2 *Avaliação dos preços de imóveis na cidade de fortaleza, com a utilização de redes neurais artificiais, para a composição do ITBI*

O estudo propõe a fazer uma avaliação dos preços na cidade de Fortaleza, com o apoio de técnicas de Redes Neurais Artificiais, para a composição do ITBI (CODES, 2018). O autor do estudo usou o software WEKA (*Waikato Environment for Knowledge Analyses*) desenvolvido pela universidade de Waikato, na Nova Zelândia. Para a calibração do modelo foram também usados os dados fornecidos pela SEFIN, com as mesmas variáveis originais desta defesa, entretanto a análise foi restrita a imóveis do tipo "Apartamento".

O autor analisou os dados para os anos de 2014, 2015 e 2016 e, em seguida. escolheu as seguintes variáveis:

1. Ano de exercício da transação;
2. Bairro;
3. Data de construção do imóvel;
4. Tipo de imóvel: favela, predial ou territorial;
5. Classificação arquitetônica: apartamento, casa, sala, loja, galpão, etc.;
6. Uso específico: residencial, comercial, industrial, etc.;
7. Ocupação: edificação, em construção, estacionamento, etc.;

8. Número de pavimentos;
9. Situação do lote: normal, esquina, gleba, etc.;
10. Fator da edificação;
11. Padrão da edificação;
12. Área do terreno;
13. Área edificada;
14. Valor de base de cálculo do ITBI;

Além do filtro de data, o autor também fez clusterização das observações por bairro, assim criou modelos diferentes apenas para os 4 bairros com maior número de observações. O índice MAPE foi usado como indicador de performance. Obtendo os seguintes resultados:

1. Meireles: 6,52%;
2. Aldeota: 6,40%;
3. Cocó: 5,44%;
4. Messejana: 4,68%.

Concluímos então que o trabalho resultou em modelos aplicáveis a esses bairros, visto que para todos foi observado um erro menor do que 15%.

3 DADOS E MÉTODOS

A SEFIN usa os dados de imóveis para o cálculo do ITBI, por conta disso, torna-se um ator público bastante interessado na formação de valores imobiliários. Dada essa realidade, o órgão mantém o histórico das operações concluídas do município e características de cada um desses bens. Os dados usados na análise contêm informações sobre transações do ano de 2009 até o ano de 2016. Destas foram usadas para a amostragem as correspondentes dos 5 últimos anos (2011 a 2015), num total de 29.502 observações.

A pesquisa será dividida em 6 etapas: Tratamento de dados; incluindo escolha de variáveis; simulação de dados via regressão linear descrito na seção 2.2.1; simulação de dados via RF disposto em 2.3.2.1; regressão gaussiana descrito em 2.2.5; GBM retratado em 2.3.3 e Redes Neurais Artificiais apresentadas em 2.2.4.

3.1 Tratamento de dados

3.1.1 Pré-Tratamento da amostra

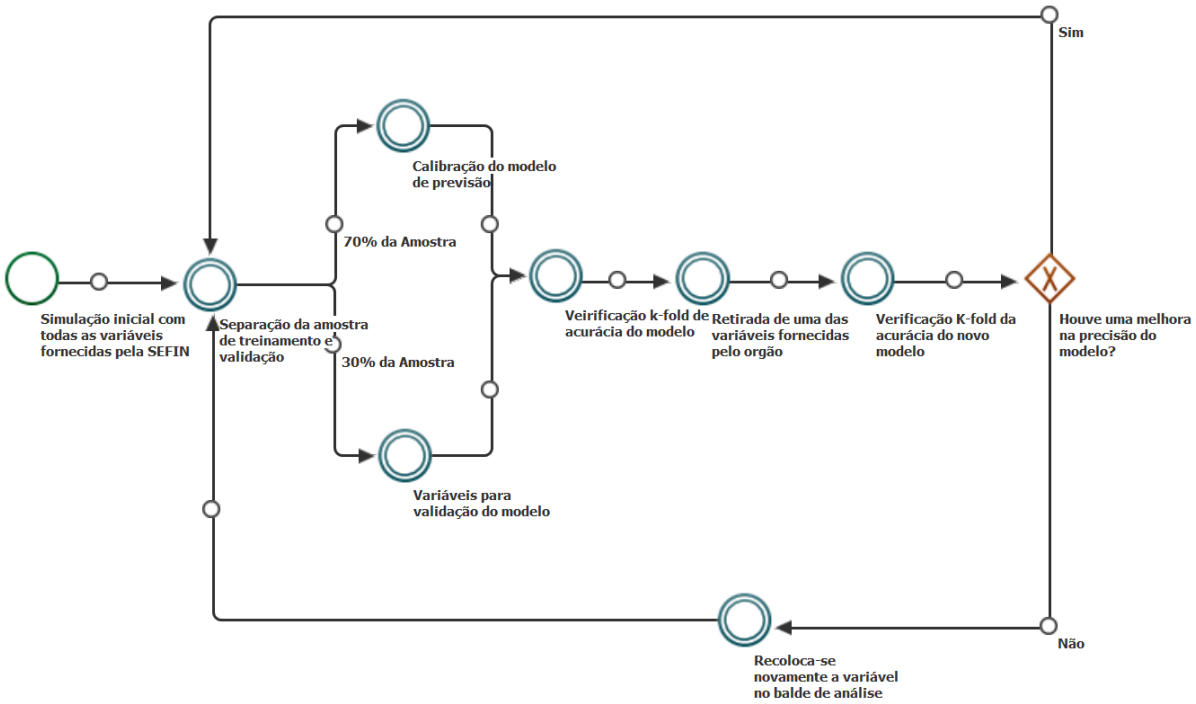
Separadas as amostras no intervalo de 2011 a 2016, será verificada a dispersão dos dados e a presença ou não de *outliers* nas observações em relação ao valor base de cálculo usado para o ITBI, Na condição de existência de *outliers*, os mesmos serão removidos da análise e após isso será verificada a dispersão dos valores de cálculo e a necessidade de ser feita alguma transformação linear, como normalização ou conversão para a escala logarítmica.

3.1.2 Escolha de Variáveis

Para garantir uma melhor aderência do modelo precisamos escolher as mais representativas para o valor do imóvel. Conforme já previsto pela (ABNT, 2011) e por ser considerado como um método de aprendizagem forte, será usada como padrão a metodologia de RNA para verificação de variáveis. Inicialmente será retirada uma sub amostra de 70% retirada aleatoriamente das observações para treinamento do modelo, sendo esta composta por todas as variáveis descritas no início do capítulo 3 para uma simulação e com ela será feita a simulação de comportamento preditivo, comparando o resultado do modelo sobre os 30% restantes da amostra e verificando sua divergência do valor real:

.

Figura 5 – Modelo de um classificador *Escolha de variáveis adotadas*



Fonte: Elaborado pelo próprio autor, 2019.

1. Verificação da precisão inicial do modelo via descrito 2.4.1.
2. Retirada de uma das variáveis fornecidas.
3. Verificação da nova precisão do modelo usando metodologia *k-Fold*.

Caso exista uma melhora no desempenho, descartamos o dado e continuamos as simulações, e em caso de piora, colocamos o dado novamente e passamos à próxima verificação. O processo irá se repetir até a verificação completa de todos os campos descritivos de transação (modelo *Stepwise*).

3.1.3 Variáveis Escolhidas

Usando a metodologia descrita na seção 3.1.2, efetuou-se simulações com todas as variáveis fornecidas e, a partir destas, escolhemos as seguintes variáveis:

Excetuando-se o modelo de regressão linear, onde apenas as variáveis quantitativas foram adotadas, nos outros modelos usamos tanto estas quanto as qualitativas, visto que lidavam bem com ambos os tipos de entradas.

Quadro 1 – Variáveis escolhidas para calibração dos modelos

Variável	Tipo	
1. Ano de exercício da Transação	Quantitativa	Discreta
2. Tipo de Logradouro - Rua, avenida, beco, estrada e etc.	Qualitativa	
3. Bairro	Qualitativa	
4. Ano da Construção	Quantitativa	Discreta
5. Uso Especifico - Agricultura, Residencial, Do Lazer e etc.	Qualitativa	
6. Ocupação - Edificação, Construção Paralisada, em construção e etc.	Qualitativa	
7. Número de Pavimentos	Quantitativa	Discreta
8. Número de Frentes	Quantitativa	Discreta
9. Número de Unidades no Lote	Quantitativa	Discreta
10. Testada Principal	Quantitativa	Continua
11. Situação do Lote com IPTU - Gleba, Esquina, Vila, Variável Quadra e etc	Qualitativa	
12. Fator Edificação	Quantitativa	Continua
13. Padrão Edificação	Qualitativa	
14. Fator Lote	Quantitativa	Continua
15. Área Terreno	Quantitativa	Continua
16. Área de Preservação	Quantitativa	Continua
17. Fração Ideal	Quantitativa	Continua
18. Área Edificada	Quantitativa	Continua
19. Valor Venal IPTU No Exercício	Quantitativa	Continua
20. Valor Base Calculo para cálculo do ITBI	Quantitativa	Continua
21. Coordenada X	Quantitativa	Continua
22. Coordenada Y	Quantitativa	Continua

Fonte: elaborado pelo autor, 2019.

3.2 Teste comparativo de modelos

Com as variáveis escolhidas, serão testadas as outras metodologias *machine-learning* descritas no capítulo 2. Os módulos serão testados individualmente e terão sua precisão aferida. Como o estudo tem o objetivo de verificar o melhor modelo, a precisão de cada preditor torna-se a variável mais importante de escolha.

Respeitando, acima de tudo, o limite estabelecido em norma de 15% de erro máximo, escolhemos aquele com menor taxa de erro percentual média e o tomamos como modelo válido para a avaliação na região. Além disso, será apresentada uma comparação entre os modelos e o realizado.

3.2.1 Hiperparâmetros

Para garantir a reprodutibilidade e acurácia do modelo, foram necessárias algumas medidas de controle, a essas sub-funções dentro do nosso modelo chamaremos de hiperparâmetros.

Os parâmetros mais básicos e compartilhados por todos os modelos são os de *seed* e o de *Number of Folds* (nfolds). O Parâmetro *seed* foi fixado em 100000 e apenas garante que os

modelos, que têm um caráter aleatório, partam sempre da mesma distribuição e obtenham sempre os mesmos resultados, garantindo uma melhoria real na troca dos demais hiperparâmetros. O parâmetro nfolds foi tomado como 10 e representa a primeira atmosfera de validação dos modelos, cada valor será validado 10 vezes pela metodologia e só então será comparado. Ademais, temos alguns hiperparâmetros específicos:

1. Parâmetros modelo de regressão Linear: Para além dos parâmetros, foram removidas as variáveis classificatórias das variáveis de treinamento. Outro fator adotado fora a troca de todas as observações faltantes (Ou do tipo *NA'S*) pelo valor médio das outras observações presentes.
2. Parâmetros modelo de regressão Gaussiana: Foi adotado o solucionador *Iteratively Reweighted Least Squares Method* (IRLSM), que é um modelo de regressão que iterativamente tenta diminuir a soma dos quadrados das diferenças entre os valores estimados e previstos. Variáveis categóricas ausentes são agrupadas na classe *missing*, já as numéricas serão substituídas pela média dos valores na categoria.
3. Parâmetros modelo *Random Forests*: Escolhemos o histograma do tipo *"UniformAdaptive"*, nesse gênero de simulação o tipo do histograma afetará apenas as variáveis continuas, nesse caso especifico as dividirá em intervalos fixos iguais para serem aplicados nas árvores de decisão. Além disso, estabelecemos 300 como número de árvores simuladas para avaliação, a partir do qual empiricamente não houve aumento de precisão com aumento do número de árvores. Variáveis categóricas e numéricas ausentes são agrupadas na classe *missing*.
4. Parâmetros modelo de Redes Neurais Artificiais: Foi realizada a normalização de todos os valores, melhorando consideravelmente a capacidade de previsão do modelo. Assim como o modelo de regressão linear, também trataremos os valores do tipo *NA's* pelo valor médio da coluna (para variáveis categóricas ausentes foi criada uma nova categoria *missing* e as entradas sem valor foram alocadas nessa categoria). O tipo de distribuição gaussiana foi adotado como previsão de entrada A a estrutura da rede neural foi criada através de duas camadas ocultas de 200 neurônios cada. Por fim, foi dito que a função perda poderia ser representada por uma função quadrática e foi admitido como 10 o número de épocas, ou retroalimentações, para a rede, valor a partir do qual não houve aumento na precisão total do preditor.
5. Parâmetros modelo de GBM: Assim como as *Random Forests* descritas no tópico 3,

adotamos um número de árvores igual a 300. Entretanto, sobre o comportamento da distribuição de entrada, graças a função gradiente será possível otimizar o resultado implicando que o histograma de entrada se assemelha a uma gaussiana. A ausência de variáveis categóricas ou numéricas será levada em consideração e tratada como uma categoria separada para a análise.

3.2.1.1 Droupout

Um dos grandes dilemas ao se estabelecer uma rede neural multi-perceptron se dá na escolha da quantidade de neurônios nas camadas intermediárias, pela dificuldade de definir um modelo ótimo de escolha para esse tipo de elemento. Embora existam algumas abordagens, o problema ainda contínua em aberto (BARBOSA, 2004).

Para modelos com poucos nós temos uma redução na precisão, embora o modelo aumente o seu índice de previsão geral. Entretanto, quando se trabalha com um número grande de neurônios, podemos ter um fenômeno chamado de *overfit*, quando o modelo acaba "decorando"por treinar demais e perde sua capacidade de previsão.

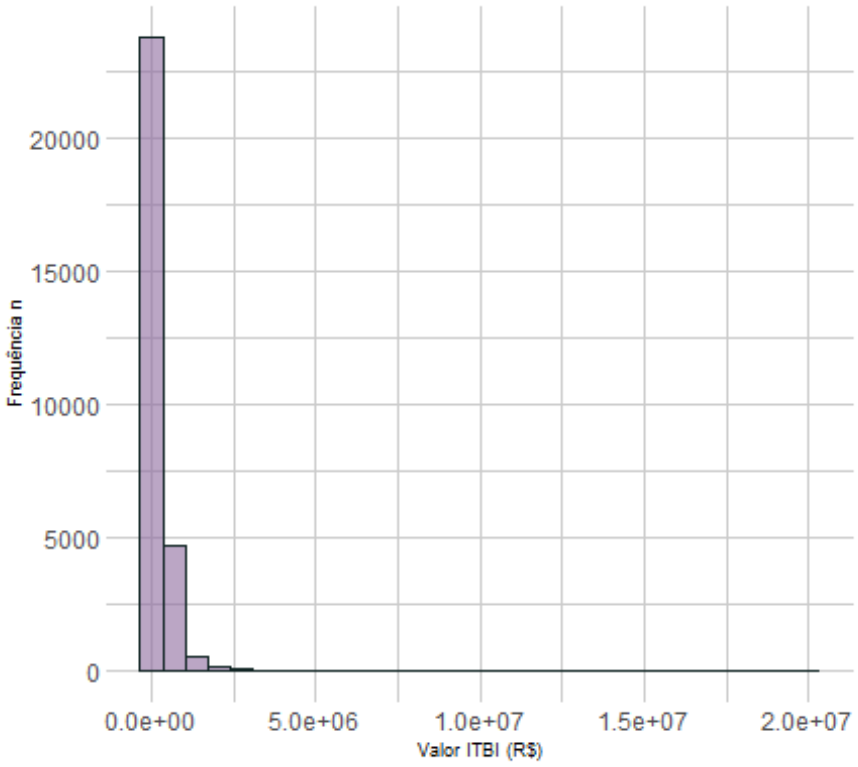
A quantidade inicial de nós é heuristicamente escolhida e em seguida é verificada a performance do modelo. Em seguida, na hipótese de uma boa varia-se o número de nós do oráculo: Quando iniciamos com uma *MLP* pequena e o número for incapaz de suprir as necessidades do projeto, aumenta-se o número total de nós, já quando a quantidade de nós for grande e se tenha uma baixa acurácia, reduzimos o número total de neurônios. A esse processo denominamos de poda (ou *prunning*).

Alguns algoritmos possuem ótimos métodos que evitam o *overfit*, o módulo H2O possui uma técnica avançada chamada de *dropout* que randomiza e desconsidera aleatoriamente alguns nós em cada simulação no modelo de previsão e, dadas as novas ponderações, temos um modelo que sempre tentará se afastar do *overfit*.

4 RESULTADOS E DISCUSSÕES

O primeiro segmento de análise foi verificar a disposição das observações para avaliar seu comportamento. Verificando o histograma de valores, vemos que se torna difícil fazer uma análise dos dados não processados, dada sua granularidade.

Figura 6 – Histograma em escala Real dos dados SEFIN



Fonte: Elaborado pelo próprio autor, 2019.

.

Ao ser simulado em um modelo de Rede Neural Artificial, obteve-se baixa acurácia na previsão. Um modelo de RNA com os valores brutos resultou em um erro médio quadrático de 104349352311 reais, medida esta que avaliaria o modelo como impraticável para o mercado.

Tabela 1 – Desempenho modelo para dados sem pré-processamento

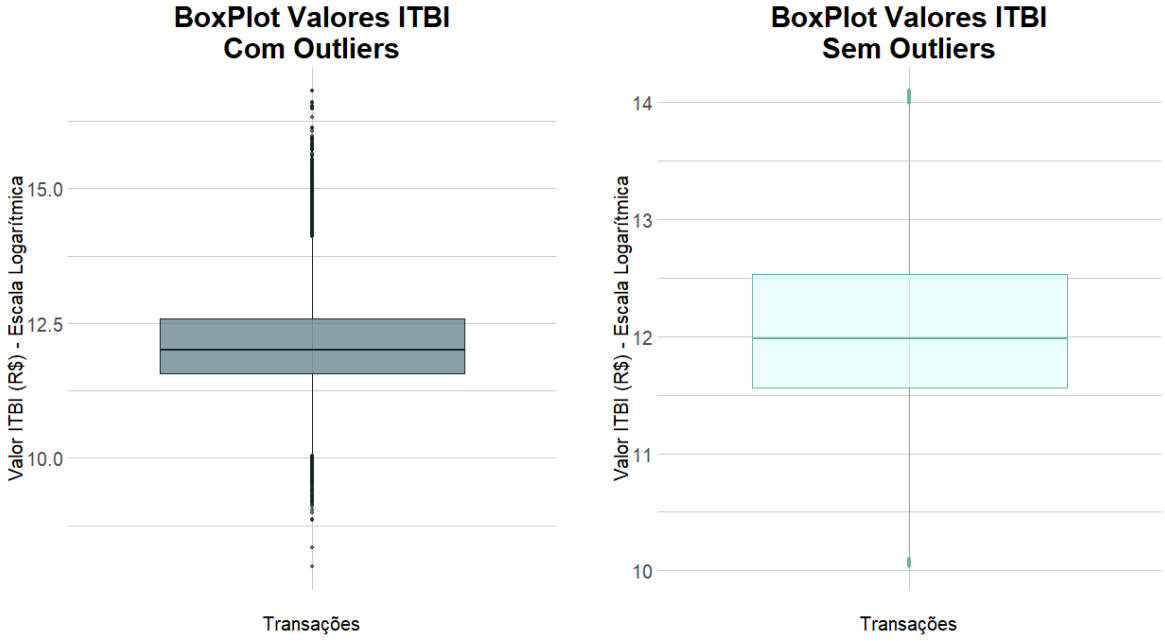
Medida	índice
MSE:	104349352311
RMSE:	323031.5
MAE:	102507.4
RMSLE:	NaN
Mean Residual Deviance :	104349352311

Fonte: o Autor, 2019.

Após converter o VALOR_BASE_CALCULO_ITBI para a escala logarítmica, conseguimos, conforme a Figura 8, obter um comportamento no histograma mais próximo a uma gaussiana, e este será o ponto base para tratarmos os modelos.

Outro fator importante para a amostragem pré modelagem fora eliminação dos *outliers* presentes em relação a variável de análise, deixando as variáveis mais uniformes para os modelos de regressão.

Figura 7 – Gráfico *box_plot* com e sem *outliers*



Fonte: Elaborado pelo próprio autor, 2019.

.

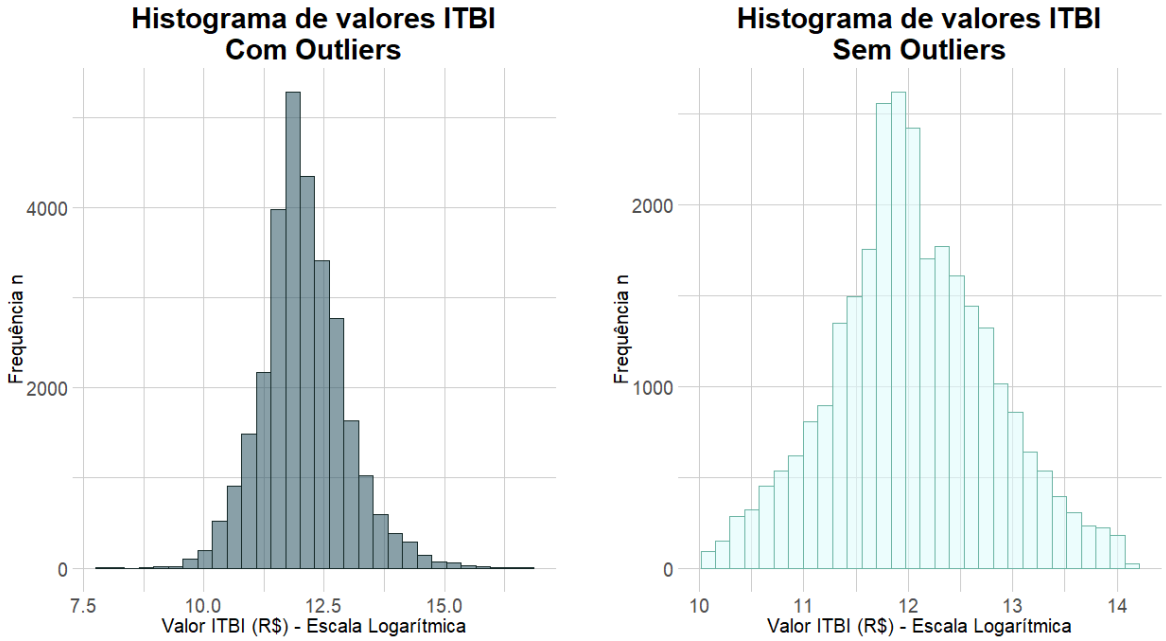
.

.

Ao todo foram removidos 878 observações do tipo *outlier*, o que representa 2,97% da amostra. Essa baixa presença de *outliers* retirados na amostra ajuda a reforçar a generalização do modelo para o tipo de imóvel analisado. Visto que os preditores representarão a grande maioria de unidades habitacionais a serem avaliadas.

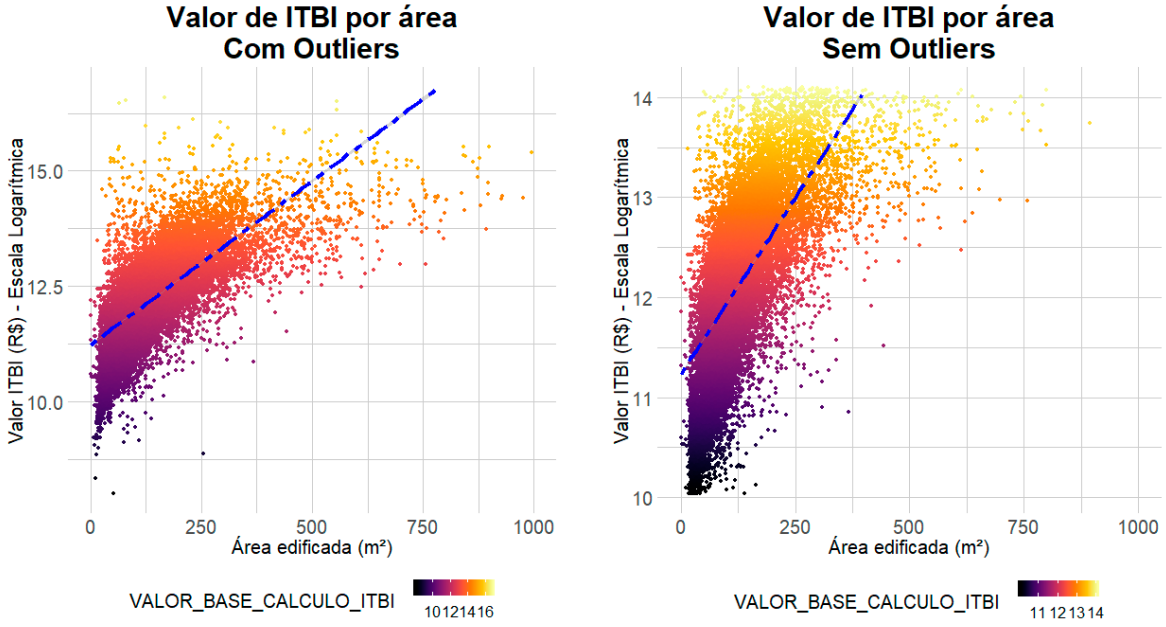
Em sequência apresentam-se os modelos usados para a avaliação usando os dados tratados. Nas páginas de apêndice serão disponibilizadas as relações de importância de variáveis para cada uma das avaliações.

Figura 8 – Histograma VALOR_BASE_CALCULO_ITBI
em escala logarítmica com e sem outliers



Fonte: Elaborado pelo próprio autor, 2019.

Figura 9 – Dados VALOR_BASE_CALCULO_ITBI/Área com e sem *outliers*



Fonte: Elaborado pelo próprio autor, 2019.

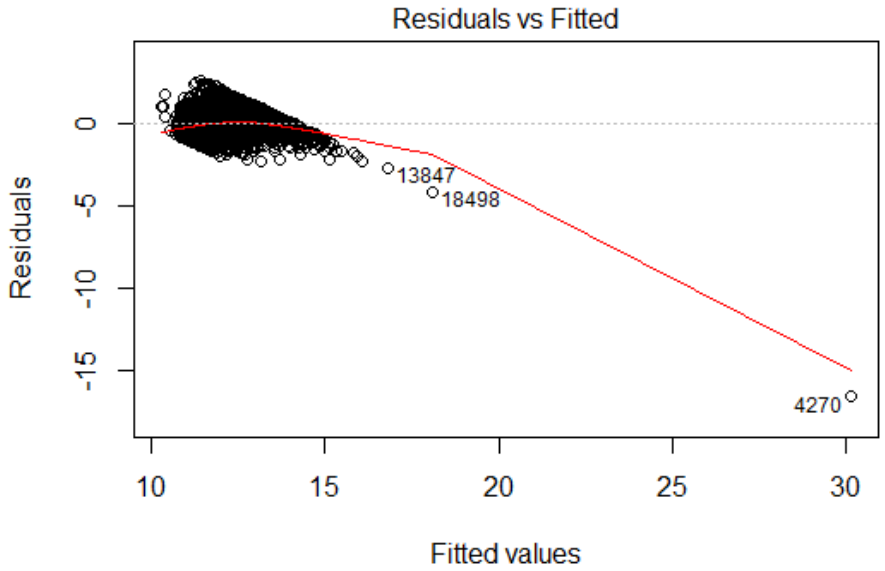
4.1 Regressão Linear

O modelo de regressão linear embora simples, obteve uma grande precisão na análise e torna-se ainda melhor trabalhando em um universo com um vasto número de observações,

por isso é um dos mais aconselhados pela (ABNT, 2011) para a análise imobiliária. Entretanto, devem-se tomar ressalvas, reavaliando o modelo para cada mudança grande no banco de dados de entrada para garantir a linearidade das observações.

O que nos traz no final a um erro médio percentual de 2,85%, valor que torna o modelo por si só indicado para a avaliação final.

Figura 10 – Valores residuais e valores ajustados



Fonte: Elaborado pelo próprio autor, 2019.

.

Tabela 2 – Desempenho modelo de regressão linear

Medida	indices
MSE:	0.3296475
RMSE:	0.57414937
MAE:	0.34012159
RMSLE:	0.03766742

Fonte: o Autor, 2019.

4.2 Regressão a Gaussiana

Devido à proximidade do histograma ajustado com uma curva normal, apresentado na Figura 8, usou-se o modelo de generalização linear com aproximação gaussiana para que se fizesse a previsão. Essa abordagem resultou em:

Tabela 3 – Desempenho modelo de regressão gaussiana

Medida	índice
MSE:	0.3021684
RMSE:	0.5496984
MAE:	0.3025599
RMSLE:	0.03491752
Mean Residual Deviance:	0.3021684

Fonte: o Autor, 2019.

Embora não tão comum, nota-se, todavia, que este modelo consegue obter um erro menor que o descrito na seção 4.1 para as variáveis de performance. Além disso, possui um erro médio percentual de 2,51%. Entretanto, assim como a premissa para regressão linear deve-se verificar e tratar os dados brutos para que seja garantido o comportamento da amostra, só que dessa vez quanto a sua normalidade. Especialmente para amostras pequenas pode apresentar dificuldades em fornecer boas previsões, sendo nesses casos possível de atrelar a regressão a métodos de replicação de amostra como os descritos nas seções 2.2.2 e 2.2.3

4.3 Random Forests

O modelo não supervisionado de *Random Forests*. ainda não normatizado, tem seu caráter avaliativo especializado em modelos de classificação, entretanto a criação de árvores aleatórias torna-se também uma medida bem robusta para modelos de regressão. Para os dados da cidade de Fortaleza obtivemos os índices disponíveis na tabela 4.

Tabela 4 – Desempenho modelo de *Random Forests*

Medida	índice
MSE:	0.08357083
RMSE:	0.2890862
MAE:	0.1960285
RMSLE:	0.02241898
Mean Residual Deviance:	0.08357083
R^2:	0.8612701

Fonte: o Autor, 2019.

A capacidade do modelo de *machine-learning* de árvores de decisão com várias árvores apresentou uma capacidade avaliativa melhor do que os modelos de regressão simples e possui um erro médio percentual de 1,64%. Além disso, mostrou uma boa adequação por apresentar um fator R^2 de 86,13%

4.4 Redes Neurais Artificiais

O modelo de RNA passou a ser adotado na (ABNT, 2011) como uma metodologia avançada de avaliação e é capaz de se adaptar a diversos tipos de variáveis através de uma retroalimentação, se tornando extremamente conhecido como um modelo de previsão. Para os dados da prefeitura de Fortaleza obtivemos:

Tabela 5 – Desempenho modelo de Redes Neurais

Medida	índice
MSE:	0.08903171
RMSE:	0.2983818
MAE:	0.2124046
RMSLE:	0.02316271
Mean Residual Deviance:	0.08903171

Fonte: o Autor, 2019.

Embora seja capaz de se adaptar para diversas amostras, uma das desvantagens do modelo é sua estrutura de *black-box* o que impossibilita que se possa verificar como as variáveis estão interagindo dentro do modelo. Embora recomendado pela norma o uso de algum algoritmo de *bagging*, não se tornou necessário devido ao tamanho da amostra, que por sua vez foi suficiente para garantir uma boa previsão e um erro médio percentual de 1,77%.

4.5 Gradient Boosting Machine

O modelo de GBM. surgiu como uma melhoria do modelo de *Random Forests*. combinando a técnica das árvores aleatórias com o princípio da aprendizagem fraca do *boosting*, descrito na seção 2.2.3. Assim como o próprio modelo de *Random Forests*, descrito em 4.3, o modelo se apresenta como um forte candidato preditor, com medidas parecidas e próximo valor R^2 mas um pouco maior (86,32%) e resultando em um erro médio percentual de 1,65%.

Tabela 6 – Desempenho modelo de *Gradient Boosting Machine*

Medida	indices
MSE:	0.08086704
RMSE:	0.2843713
MAE:	0.1971285
RMSLE:	0.02211454
Mean Residual Deviance:	0.08086704
R^2:	0.8632644

Fonte: o Autor, 2019.

4.6 Discussões sobre os modelos

Tabela 7 – Resultados Comparativo de erros médios percentuais modelados

Medida	MAPE (%)
Regressão Linear Múltipla (RLM)	2.85
Regressão Gaussiana (GLM)	2.51
RNA	1.77
RF	1.64
GBM	1.65

Fonte: o Autor, 2019.

Notou-se que variáveis classificatórias possuem um papel significativo para os oráculos, sendo que para o modelo de regressão Gaussiana, sem as variáveis classificatórias obtivemos um erro médio percentual de 2,85%, com essas variáveis obtemos um erro de 2,51%, uma diferença percentual de 11% entre as duas previsões.

Cabe acrescentar também que para modelos de caráter predominantemente classificatório como RF e GBM obtivemos os melhores resultados, o que reforça a hipótese da força das variáveis classificatórias.

Além disso, o pré-processamento de dados é muito importante, medidas como a remoção de *outliers* e a normatização do valor para cálculo de ITBI representaram uma diferença entre MSE de 104349352311 com amostra não tratada para um MSE de 0.0933914 pós tratamento, uma diferença de mais de 99,99%, isto se deve principalmente a diferença de preços nas transações imobiliárias, na nossa base não tratada haviam valores para ITBI que variavam de R$ 2983,09 até valores da ordem de 20 milhões de reais. O comparativo entre valores MAE e RMSE ilustra a influência da presença *outliers* na amostra, dado que este último índice é mais volátil a presença deles. Portanto, devido sua proximidade, foi possível obter um

modelo consistente para valores entre o primeiro e o terceiro quartil, criando um modelo mais adequado para a maioria das transações.

5 CONCLUSÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Com o término desse trabalho, foi possível notar a forte capacidade preditiva de modelos estatísticos para a previsão do preço de imóveis. Um fator determinante para a qualidade das previsões foi, sem dúvida, o tamanho da amostragem, pois, graças a quantidade de observações, foi possível atingir uma distribuição próxima a normalidade, onde mesmo modelos mais simples como o de regressão linear tiveram um resultado com ordem de erro abaixo dos 3%. Por conta disso, recomenda-se aos profissionais de avaliação que trabalhem com modelos mais generalistas, evitando a criação de modelos diversos para determinados grupamentos de amostra.

Verificando a influência das variáveis disponíveis nos Apêndices B, C, D, E e F é possível notar uma tendência maior de que a localidade onde está o imóvel influencia o preço final de venda. Para além do padrão de casa, determinados bairros tendem a influenciar mais no preço final. Em seguida, aparecem variáveis de dimensões da propriedade, como área ou testada, indicando que o consumidor teria uma tendência maior em escolher um imóvel por suas características que tendem a uma maior imutabilidade.

Todos os modelos simulados são adequados para a previsão proposta no trabalho, visto que todos possuem um erro significativamente abaixo dos 15%, entretanto os modelos de aprendizagem assistida apresentaram melhores índices de previsão, sendo a diferença do modelo de regressão linear sem as variáveis categóricas para o modelo mais preciso de *machine-learning* de 26,22%, valor expressivo que pode significar ou não a validação do modelo para a amostra desejada.

As principais dificuldades desse trabalho foram frente ao processamento de dados base, estabelecer os procedimentos para o tratamento de dados base foi de fundamental importância porque devido a falta de uniformidade nos valores de ITBI, mesmo os modelos mais robustos não conseguiram obter exito em processar os dados brutos. Outro desafio se baseou em cima do processo de calibragem de modelos, para alguns parâmetros, como o número de árvores para os modelos de RF e GBM, foi necessário realizar uma análise empírica para determinar melhor o valor usado. Além disso, conforme análise feita no apêndice A, cada um dos modelos tem suas vantagens, desvantagens e particularidades, e estas devem ser levadas em consideração durante o processo de escolha de modelos base

Para esse estudo, desejou-se apenas obter a previsão do preço dos imóveis do tipo "Casa" na cidade de Fortaleza. Para estudos complementares recomenda-se a avaliação dos modelos para outros tipos de edificação, como apartamentos, lojas, salas, galpões e entre outros,

visto que os mesmos podem apresentar comportamentos diferentes durante a previsão. Outro estudo complementar possível poderia recriar os modelos acrescidos de novas variáveis de localização, como proximidade a centros comerciais, a delegacias, a hospitais e a ambientes de lazer. Uma grande aliada da aprendizagem de máquina atualmente é a área da geoestatística, o uso das medidas de dispersão espacial ou o emprego da krigagem para a verificação dos comportamentos, podem ser grandes aliados durante o processo de avaliação imobiliária.

Por fim, através do mecanismo mais preciso, recomenda-se o uso das metodologias RF ou GBM para a previsão dos imóveis vendidos na cidade, pois apresentaram o menor erro percentual na análise e não apresentam diferenças significativas entre os resultados obtidos nelas. Embora RF tenha sido mais preciso, deve-se sempre averiguar a GBM para novas simulações, visto que a função gradiente que nela é atrelada a torna mais resistente a alterações significativas de base.

Concluímos então que o estudo atingiu seus objetivos, os modelos de aprendizagem de máquina podem ser usados no âmbito das avaliações imobiliárias quando devidamente aplicados. Também concluímos que podem ser metodologias de igual acurácia ou até superiores aos modelos de regressão linear. Por conta desse fator, além do mecanismo de RNA, já descrito em norma, o autor recomenda a avaliação e a exposição em norma de mais mecanismos avançados de avaliação imobiliária, como os descritos nesse trabalho. Assim, obteremos previsões mais assertivas e melhores transações no mercado.

Por fim, através do mecanismo mais preciso, recomenda-se o uso das metodologias RF ou GBM para a previsão dos imóveis vendidos na cidade, pois apresentaram o menor erro percentual na análise e não apresentam diferenças significativas entre os resultados obtidos nelas. Embora RF tenha sido mais preciso, deve-se sempre averiguar a GBM para novas simulações, visto que a função

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653**: Avaliação de bens parte 2: Imóveis urbanos. 2. ed. Rio de Janeiro, 2011. 62 p.

BARBOSA, A. T. R. Mecanismo de Adaptação Baseado Em Redes Neurais Artificiais Para Sistemas Hipermídia Adaptativos. p. 123, 2004.

BISHOP, C. **Neural networks for pattern recognition**. Oxford New York: Clarendon Press Oxford University Press, 1995. ISBN 978-0198538646.

BREIMAN, L. Random Forests. transparencias. **Statistics**, v. 45, n. 1, p. 1–33, 2001. ISSN 08856125. Disponível em: <http://dx.doi.org/10.1023/A:1010933404324>.

CHEN, Y. 2017. **A collection of evaluation metrics, including loss, score and utility functions, that measure regression, classification and ranking performance**). Disponível em: <https://github.com/yanyachen/MLmetrics>. Acesso em: 20 out. 2019.

CODES, B. N. **Modelo De Redes Neurais Artificiais Para Avaliar a Formação De Preços De Imóveis Na Cidade De Fortaleza**. [S.l.], 2018. 82 p.

DIPASQUALE, D.; WHEATON, W. C. **Urban economics and real estate markets**. [S.l.]: Prentice Hall Englewood Cliffs, NJ, 1996. v. 23.

ELDIONARA, M.; MESTRANDA, R. M.; ADMINISTRAÇÃO, E.; CERETTA, P. S.; VIEIRA, K. M. a Relação Entre As Variáveis Macroeconômicas E a Concessão De Crédito No Mercado Imobiliário Brasileiro the Relationship Between the Macroeconomic Variables and the Granting of Credit in the Real Estate Market Brazilian La Relación Entre Variables Macroec. p. 64–84, 2014.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **Regularization Paths for Generalized Linear Models via Coordinate Descent**. [S.l.], 2010. v. 33. Disponível em: <http://www.jstatsoft.org/>.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001. ISSN 00905364. Disponível em: <http://www.jstor.org/stable/2699986>.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C.; VILLA-VIALANEIX, N. Random Forests for Big Data. **Big Data Research**, Elsevier, v. 9, p. 28–46, sep 2017. ISSN 2214-5796. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214579616301939{\#}fg0>.

H2O.AI. 2019. **The metrics for this section only cover supervised learning models, which vary based on the model type (classification or regression**). Disponível em: <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/performance-and-prediction.html>. Acesso em: 04 out. 2019.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 1st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1994. ISBN 0023527617.

KOHAVI, R. Cross validation number.pdf. v. 5, p. 7, 1995. ISSN 10450823. Disponível em: <http://robotics.stanford.edu/{~}ronn>.

Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, Nov 1998. ISSN 0018-9219.

LEO, B. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1007/BF00058655>.

MALERE, J.; ALMEIDA, P.; SANO, H. Predição de preços de imóveis através de aprendizagem de máquina. 07 2019.

PASQUALIN, L. L. B. Os distratos no mercado imobiliário de empreendimentos residenciais de São Paulo : uma discussão sobre as situações que favorecem a interrupção de contratos. 2016.

RIBEIRO, I. P.; BERTRAN, M. P. C. Crise imobiliária brasileira: a transferência de renda pelos "distratos" e créditos podres. **Revista Eletrônica Direito e Sociedade - REDES**, v. 7, n. 1, p. 139, 2019.

SCHAPIRE, R. E. The Strength of Weak Learnability (Extended Abstract). **Machine learning**, v. 5, p. 197–227, 1990. Disponível em: <http://link.springer.com/article/10.1007/BF00116037[AccessedonNov2018]>.

SELIM, H. Determinants of house prices in turkey: Hedonic regression versus artificial neural network. **Expert Syst. Appl.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 36, n. 2, p. 2843–2852, mar. 2009. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2008.01.044>.

TSAI, C. F.; HSU, Y. F.; YEN, D. C. A comparative study of classifier ensembles for bankruptcy prediction. **Applied Soft Computing Journal**, v. 24, p. 977–984, 2014. ISSN 15684946.

APÊNDICE A – COMPARATIVO ENTRE OS MODELOS ABORDADOS NO TRABALHO

Modelos	Vantagens	Desvantagens
Regressão Linear Múltipla	- Fácil de inspeção de comportamentos de variáveis pós treinamento - Modelo não necessita de grande capacidade computacional para realizar treinamento - Fácil de calibragem - Modelo já definido na norma de avaliações imobiliárias	- Apenas adequado para modelos da regressão - Dificuldade com amostras muito pequenas
Random Forests	- Viável de ser usado para Regressão e Classificação - Modelo de fácil calibração - Trabalha com variáveis Quantitativas e Qualitativas	- Menos acurado para amostras pequenas - Modelo próximo a uma Black-Box. Difícil de verificar comportamentos entre variáveis
GLM	- Fácil de calibragem - Trabalha com variáveis Quantitativas e Qualitativas - Modelo não necessita de grande capacidade computacional para realizar treinamento	- Necessário garantir distribuição de dados (Gaussiana, Poisson, Gamma, Ordinal e etc.) - Dificuldade em lidar com amostras muito pequenas
GBM	- Mais acurado para amostras pequenas que o modelo de *Random Forests* - Trabalha com variáveis Quantitativas e Qualitativas - Viável de ser usado para Regressão e Classificação	- Volátil a ruído nas amostras de treinamento - Mais difíceis de Calibrar que as *Random Forests* devido a função de *Boosting* - Modelo próximo a uma Black-Box. Difícil inspeção das relações entre variáveis pós modelo - Demanda grande quantidade de processamento do computador para modelo de treinamento
Redes Neurais Artificiais	- Viável de ser usado para Regressão e Classificação - Trabalha com variáveis Quantitativas e Qualitativas - Não necessária garantia de uniformidade do comportamento de variáveis para treinamento - Modelo já definido na norma de avaliações imobiliárias	- Modelo Black Box, o que torna impossível para inspeção do comportamento de variáveis pós modelo - Demanda grande quantidade de processamento do computador para modelo de treinamento - Sensível ao overfitting - Preditor de difícil calibração, necessário maior cuidado para ser aplicado

Fonte: Elaborado pelo Autor, 2019.

APÊNDICE B – IMPORTÂNCIA DE VARIÁVEIS PARA O MODELO GBM

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
VALOR.VENAL.IPTU.NO.EXERCICIO	16533,71484375	100	54,5
AREA.EDIFICADA.GI	5575,939453125	33,72	18,38
BAIRRO	3679,5146484375	22,25	12,13
DATA.CADASTRO.GI	1484,41625976562	8,98	4,89
AREA.TERRENO.GI	1339,17907714844	8,1	4,41
DATA.CONSTRUCAO	511,321533203125	3,09	1,69
FATOR.EDIFICACAO	247,715728759766	1,5	0,82
FRACAO.IDEAL	240,809341430664	1,46	0,79
PADRAO.EDIFICACAO	142,907165527344	0,86	0,47
TESTADA.PRINCIPAL	138,669525146484	0,84	0,46
Y	128,929565429688	0,78	0,42
X	85,6877593994141	0,52	0,28
FATOR.LOTE	77,8000335693359	0,47	0,26
NUMERO.PAVIMENTOS	38,2704696655273	0,23	0,13
TIPO.LOGRADOURO	37,8358612060547	0,23	0,12
NUM.UNIDADES.LOTE	22,2610034942627	0,13	0,07
NUMERO.FRENTES	18,4172973632812	0,11	0,06
USO.ESPECIFICO	17,6623630523682	0,11	0,06
SITUACAO.LOTE.IPTU	15,6752605438232	0,09	0,05
OCUPACAO	2,79046821594238	0,02	0,01

Fonte: o Autor, 2019.

APÊNDICE C – IMPORTÂNCIA DE VARIÁVEIS PARA O MODELO GLM

(Continua)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.DIONISIO TORRES	51,07	100	4,43
BAIRRO.ALDEOTA	50,91	99,68	4,42
PADRAO.EDIFICACAO.Baixo nivel 1	44,21	86,57	3,84
BAIRRO.GUARARAPES	43,91	85,97	3,81
BAIRRO.MEIRELES	37,7	73,83	3,27
BAIRRO.COCO	35,71	69,93	3,1
AREA.EDIFICADA.GI	32,68	64	2,84
BAIRRO.JANGURUSSU	32,28	63,2	2,8
BAIRRO.FATIMA	29,23	57,23	2,54
BAIRRO.VARJOTA	27,81	54,45	2,41
BAIRRO.VICENTE PINZON	27,11	53,08	2,35
TIPO.LOGRADOURO.AVENIDA	27,08	53,02	2,35
BAIRRO.JARDIM IRACEMA	26,63	52,14	2,31
BAIRRO.BARROSO	24,76	48,49	2,15
BAIRRO.BARRA DO CEARA	23,97	46,94	2,08
BAIRRO.JACARECANGA	23,56	46,14	2,05
BAIRRO.QUINTINO CUNHA	23,08	45,18	2
BAIRRO.PARQUELANDIA	22,81	44,66	1,98
PADRAO.EDIFICACAO.Normal 2	22,46	43,97	1,95
BAIRRO.CIDADE 2000	21,52	42,14	1,87

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.CIDADE DOS FUNCIONARIOS	21,35	41,8	1,85
DATA.CADASTRO.GI	20,81	40,75	1,81
X	20,55	40,24	1,78
BAIRRO.MARAPONGA	20,37	39,88	1,77
BAIRRO.JARDIM AMERICA	20,25	39,66	1,76
BAIRRO.GUAJERU	17,92	35,1	1,56
BAIRRO.LAGOA REDONDA	17,15	33,59	1,49
TIPO.LOGRADOURO.RUA	15,81	30,95	1,37
TESTADA.PRINCIPAL	15,53	30,42	1,35
BAIRRO.MONDUBIM	15,46	30,28	1,34
BAIRRO.AEROLANDIA	15,04	29,45	1,31
TIPO.LOGRADOURO.VILA	14,57	28,53	1,27
BAIRRO.PARANGABA	12,98	25,42	1,13
BAIRRO.JOSE DE ALENCAR	12,97	25,4	1,13
BAIRRO.PAUPINA	12,57	24,62	1,09
BAIRRO.PARQUE MANIBURA	11,89	23,29	1,03
Y	11,73	22,97	1,02
SITUACAO.LOTE.IPTU.Vila	11,59	22,69	1,01
BAIRRO.ENGENHEIRO LUCIANO CAVALCANTE	11,53	22,58	1
BAIRRO.CAMBEBA	11,45	22,42	0,99
BAIRRO.JARDIM GUANABARA	10,85	21,24	0,94

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.ALTO DA BALANCA	10,69	20,93	0,93
FATOR.EDIFICACAO	10,56	20,69	0,92
BAIRRO.JOAQUIM TAVORA	10,24	20,05	0,89
BAIRRO.VILA VELHA	9,43	18,46	0,82
BAIRRO.FLORESTA	8,99	17,6	0,78
DATA.CONSTRUCAO	8,68	17	0,75
BAIRRO.ANCURI	8,63	16,9	0,75
TIPO.LOGRADOURO.ALAMEDA	8,51	16,66	0,74
BAIRRO.AMADEU FURTADO	8,3	16,25	0,72
BAIRRO.PARQUE DOIS IRMAOS	7,39	14,47	0,64
BAIRRO.MONTE CASTELO	7,2	14,1	0,63
BAIRRO.SAPIRANGA	6,99	13,69	0,61
NUM.UNIDADES.LOTE	6,53	12,79	0,57
FRACAO.IDEAL	6,26	12,26	0,54
SITUACAO.LOTE.IPTU.Esquina	5,74	11,23	0,5
BAIRRO.DENDE	5,68	11,13	0,49
BAIRRO.SIQUEIRA	5,61	10,99	0,49
BAIRRO.VILA UNIAO	5,16	10,1	0,45
AREA.TERRENO.GI	3,81	7,46	0,33
BAIRRO.PASSARE	3,63	7,1	0,31
SITUACAO.LOTE.IPTU.Gleba	3,57	6,98	0,31
USO.ESPECIFICO.Comercial	3,53	6,91	0,31
BAIRRO.MONTESE	3,34	6,55	0,29

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.CENTRO	3,2	6,27	0,28
BAIRRO.MESSEJANA	3,03	5,94	0,26
BAIRRO.COACU	3	5,88	0,26
BAIRRO.FARIAS BRITO	2,77	5,42	0,24
PADRAO.EDIFICACAO.Alto nivel 1	2,43	4,75	0,21
PADRAO.EDIFICACAO.Normal 3	2,34	4,59	0,2
BAIRRO.BOM JARDIM	2,27	4,44	0,2
BAIRRO.BOA VISTA	2,21	4,33	0,19
PADRAO.EDIFICACAO.	2,17	4,25	0,19
FATOR.LOTE	2,1	4,12	0,18
BAIRRO.GRANJA LISBOA	1,98	3,87	0,17
BAIRRO.PARQUE IRA-CEMA	1,92	3,76	0,17
NUMERO.PAVIMENTOS	1,92	3,75	0,17
BAIRRO.EDSON QUEIROZ	1,83	3,59	0,16
SITUACAO.LOTE.IPTU.Normal	1,68	3,29	0,15
BAIRRO.HENRIQUE JORGE	1,53	3,01	0,13
BAIRRO.JARDIM DAS OLI-VEIRAS	1,52	2,98	0,13
BAIRRO.PARQUE SANTA MARIA	1,49	2,93	0,13
BAIRRO.CONJUNTO CEARA	0,98	1,92	0,09
AREA.PRESERVACAO.GI	0,81	1,59	0,07
NUMERO.FRENTES	0,58	1,14	0,05

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
VALOR.VENAL.IPTU.NO.EXERCICIO	0,27	0,53	0,02
BAIRRO.ALVARO WEYNE	0,24	0,46	0,02
BAIRRO.	0	0	0
BAIRRO.AEROPORTO	0	0	0
BAIRRO.ANTONIO BEZERRA	0	0	0
BAIRRO.AUTRAN NUNES	0	0	0
BAIRRO.BELA VISTA	0	0	0
BAIRRO.BENFICA	0	0	0
BAIRRO.BOM FUTURO	0	0	0
BAIRRO.BOM SUCESSO	0	0	0
BAIRRO.CAIS DO PORTO	0	0	0
BAIRRO.CAJAZEIRAS	0	0	0
BAIRRO.CANINDEZINHO	0	0	0
BAIRRO.CARLITO PAMPLONA	0	0	0
BAIRRO.CONJUNTO CEARA I	0	0	0
BAIRRO.CONJUNTO ESPERANCA	0	0	0
BAIRRO.CONJUNTO PALMEIRAS	0	0	0
BAIRRO.COUTO FERNANDES	0	0	0
BAIRRO.CRISTO REDENTOR	0	0	0
BAIRRO.CURIO	0	0	0
BAIRRO.DAMAS	0	0	0

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.DE LOURDES	0	0	0
BAIRRO.DEMOCRITO ROCHA	0	0	0
BAIRRO.DIAS MACEDO	0	0	0
BAIRRO.DOM LUSTOSA	0	0	0
BAIRRO.GENIBAU	0	0	0
BAIRRO.GRANJA PORTUGAL	0	0	0
BAIRRO.ITAOCA	0	0	0
BAIRRO.ITAPERI	0	0	0
BAIRRO.JARDIM CEARENSE	0	0	0
BAIRRO.JOSE. BONIFACIO	0	0	0
BAIRRO.JOAO XXIII	0	0	0
BAIRRO.JOQUEI CLUBE	0	0	0
BAIRRO.MANOEL DIAS BRANCO	0	0	0
BAIRRO.MANOEL SATIRO	0	0	0
BAIRRO.MOURA BRASIL	0	0	0
BAIRRO.MUCURIPE	0	0	0
BAIRRO.NOVO MONDUBIM	0	0	0
BAIRRO.OLAVO OLIVEIRA	0	0	0
BAIRRO.PADRE ANDRADE	0	0	0
BAIRRO.PAN AMERICANO	0	0	0

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.PAPICU	0	0	0
BAIRRO.PARQUE ARAXA	0	0	0
BAIRRO.PARQUE PRESIDENTE VARGAS	0	0	0
BAIRRO.PARQUE SANTA ROSA	0	0	0
BAIRRO.PARQUE SAO JOSE	0	0	0
BAIRRO.PARRAO	0	0	0
BAIRRO.PEDRAS	0	0	0
BAIRRO.PICI	0	0	0
BAIRRO.PLANALTO AYRTON SENNA	0	0	0
BAIRRO.PRAIA DE IRACEMA	0	0	0
BAIRRO.PRAIA DO FUTURO I	0	0	0
BAIRRO.PRAIA DO FUTURO II	0	0	0
BAIRRO.PREFEITO JOSE. VALTER	0	0	0
BAIRRO.PRESIDENTE KENNEDY	0	0	0
BAIRRO.RODOLFO TEOFILO	0	0	0
BAIRRO.SABIAGUABA	0	0	0
BAIRRO.SALINAS	0	0	0
BAIRRO.SERRINHA	0	0	0

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.SAO BENTO	0	0	0
BAIRRO.SAO GERARDO	0	0	0
BAIRRO.SAO JOAO DO TAUAPE	0	0	0
BAIRRO.VILA ELLERY	0	0	0
BAIRRO.VILA PERY	0	0	0
USO.ESPECIFICO.Comunicacao	0	0	0
USO.ESPECIFICO.Do lazer	0	0	0
USO.ESPECIFICO.Fechado	0	0	0
USO.ESPECIFICO.Hotelaria	0	0	0
USO.ESPECIFICO.Industrial	0	0	0
USO.ESPECIFICO.Institucional	0	0	0
USO.ESPECIFICO.Instrucao	0	0	0
USO.ESPECIFICO.Prestacao	0	0	0
USO.ESPECIFICO.Religioso	0	0	0
USO.ESPECIFICO.Residencial	0	0	0
USO.ESPECIFICO.Saude	0	0	0
USO.ESPECIFICO.Sem	0	0	0
PADRAO.EDIFICACAO.Alto nivel 2	0	0	0
PADRAO.EDIFICACAO.Alto nivel 3	0	0	0
PADRAO.EDIFICACAO.Baixo nivel 2	0	0	0
PADRAO.EDIFICACAO.Baixo nivel 3	0	0	0
PADRAO.EDIFICACAO.Luxo 1	0	0	0

Tabela 8 – Importância de variáveis para o modelo GLM

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
PADRAO.EDIFICACAO.Luxo 2	0	0	0
PADRAO.EDIFICACAO.Normal 1	0	0	0
TIPO.LOGRADOURO.BECO	0	0	0
TIPO.LOGRADOURO.ESTRADA	0	0	0
TIPO.LOGRADOURO.PRACA	0	0	0
TIPO.LOGRADOURO.RODOVIA	0	0	0
TIPO.LOGRADOURO.TRAVESSA	0	0	0
SITUACAO.LOTE.IPTU.	0	0	0
SITUACAO.LOTE.IPTU.Encravado	0	0	0
SITUACAO.LOTE.IPTU.Quadra	0	0	0
OCUPACAO.Construcao paralisada	0	0	0
OCUPACAO.Edificacao	0	0	0
OCUPACAO.Em construcao	0	0	0
OCUPACAO.Ruinas/demolicao	0	0	0
OCUPACAO.SEM	0	0	0

Fonteo Autor, 2019.

APÊNDICE D – IMPORTÂNCIA DE VARIÁVEIS PARA O MODELO RNA

(Continua)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
VALOR.VENAL.IPTU.NO.EXERCICIO	100	100	1,75
AREA.TERRENO.GI	79,1	79,1	1,39
DATA.CONSTRUCAO	69,44	69,44	1,22
AREA.EDIFICADA.GI	62,31	62,31	1,09
TESTADA.PRINCIPAL	56,44	56,44	0,99
FRACAO.IDEAL	45,93	45,93	0,81
OCUPACAO.Edificacao	41,82	41,82	0,73
BAIRRO.PASSARE	41,15	41,15	0,72
Y	39,69	39,69	0,7
NUMERO.PAVIMENTOS	38,79	38,79	0,68
AREA.PRESERVACAO.GI	38,3	38,3	0,67
DATA.CADASTRO.GI	38,06	38,06	0,67
X	37,88	37,88	0,66
USO.ESPECIFICO.Residencial	37,18	37,18	0,65
SITUACAO.LOTE.IPTU.Normal	37,07	37,07	0,65
PADRAO.EDIFICACAO.Normal 3	36,96	36,96	0,65
NUM.UNIDADES.LOTE	36,76	36,76	0,65
BAIRRO.MONDUBIM	36,16	36,16	0,63
BAIRRO.PAPICU	35,99	35,99	0,63
BAIRRO.PARQUELANDIA	35,95	35,95	0,63
TIPO.LOGRADOURO.RUA	35,89	35,89	0,63
BAIRRO.JANGURUSSU	35,55	35,55	0,62
BAIRRO.DAMAS	35,37	35,37	0,62
BAIRRO.JARDIM AME-RICA	35,32	35,32	0,62
TIPO.LOGRADOURO.AVENIDA	34,95	34,95	0,61

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
PADRAO.EDIFICACAO.Alto nivel 1	34,86	34,86	0,61
PADRAO.EDIFICACAO.Normal 2	34,79	34,79	0,61
BAIRRO.MARAPONGA	34,78	34,78	0,61
FATOR.EDIFICACAO	34,6	34,6	0,61
BAIRRO.MEIRELES	34,25	34,25	0,6
BAIRRO.VICENTE PINZON	34,2	34,2	0,6
BAIRRO.SAPIRANGA	33,87	33,87	0,59
PADRAO.EDIFICACAO.Alto nivel 2	33,84	33,84	0,59
BAIRRO.JOAQUIM TAVORA	33,73	33,73	0,59
BAIRRO.DOM LUSTOSA	33,32	33,32	0,58
BAIRRO.PARQUE SANTA MARIA	33,03	33,03	0,58
BAIRRO.SAO JOAO DO TAUAPE	33	33	0,58
BAIRRO.VARJOTA	32,81	32,81	0,58
BAIRRO.CENTRO	32,63	32,63	0,57
SITUACAO.LOTE.IPTU.Esquina	32,48	32,48	0,57
BAIRRO.COCO	32,46	32,46	0,57
BAIRRO.ENGENHEIRO LUCIANO CAVALCANTE	32,4	32,4	0,57
BAIRRO.MUCURIPE	32,38	32,38	0,57
BAIRRO.CIDADE 2000	32,21	32,21	0,57
BAIRRO.PARQUE IRACEMA	32,03	32,03	0,56

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.ALTO DA BALANCA	32,02	32,02	0,56
BAIRRO.BELA VISTA	32,02	32,02	0,56
BAIRRO.PICI	31,94	31,94	0,56
BAIRRO.JOAO XXIII	31,94	31,94	0,56
BAIRRO.BOM FUTURO	31,71	31,71	0,56
FATOR.LOTE	31,61	31,61	0,55
BAIRRO.AEROLANDIA	31,56	31,56	0,55
BAIRRO.JARDIM CEARENSE	31,54	31,54	0,55
BAIRRO.JOSE DE ALENCAR	31,51	31,51	0,55
BAIRRO.VILA UNIAO	31,42	31,42	0,55
BAIRRO.PARQUE MANIBURA	31,42	31,42	0,55
BAIRRO.BOM SUCESSO	31,4	31,4	0,55
BAIRRO.JARDIM DAS OLIVEIRAS	31,39	31,39	0,55
BAIRRO.PARQUE ARAXA	31,34	31,34	0,55
BAIRRO.DIAS MACEDO	31,21	31,21	0,55
BAIRRO.HENRIQUE JORGE	31,06	31,06	0,55
BAIRRO.ANCURI	30,92	30,92	0,54
BAIRRO.GUAJERU	30,84	30,84	0,54
BAIRRO.GRANJA LISBOA	30,77	30,77	0,54
BAIRRO.JACARECANGA	30,76	30,76	0,54
BAIRRO.EDSON QUEIROZ	30,76	30,76	0,54

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.MONTE CASTELO	30,73	30,73	0,54
BAIRRO.BENFICA	30,62	30,62	0,54
BAIRRO.ALDEOTA	30,62	30,62	0,54
BAIRRO.SAO GERARDO	30,6	30,6	0,54
BAIRRO.GUARARAPES	30,6	30,6	0,54
OCUPACAO.Em construcao	30,59	30,59	0,54
BAIRRO.ANTONIO BEZERRA	30,56	30,56	0,54
BAIRRO.AMADEU FURTADO	30,54	30,54	0,54
USO.ESPECIFICO.Comercial	30,54	30,54	0,54
BAIRRO.BOA VISTA	30,54	30,54	0,54
BAIRRO.LAGOA REDONDA	30,44	30,44	0,53
BAIRRO.ITAPERI	30,44	30,44	0,53
BAIRRO.QUINTINO CUNHA	30,43	30,43	0,53
BAIRRO.PARQUE DOIS IRMAOS	30,39	30,39	0,53
BAIRRO.MONTESE	30,35	30,35	0,53
BAIRRO.ITAOCA	30,27	30,27	0,53
BAIRRO.SALINAS	30,25	30,25	0,53
BAIRRO.PADRE ANDRADE	30,18	30,18	0,53
BAIRRO.SERRINHA	30,18	30,18	0,53
BAIRRO.BOM JARDIM	30,17	30,17	0,53

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.DIONISIO TORRES	30,04	30,04	0,53
BAIRRO.DENDE	29,97	29,97	0,53
PADRAO.EDIFICACAO.	29,9	29,9	0,52
BAIRRO.PLANALTO AYRTON SENNA	29,8	29,8	0,52
BAIRRO.JOSE BONIFACIO	29,79	29,79	0,52
BAIRRO.PEDRAS	29,78	29,78	0,52
TIPO.LOGRADOURO.ALAMEDA	29,76	29,76	0,52
BAIRRO.PRESIDENTE KENNEDY	29,75	29,75	0,52
BAIRRO.CAMBEBA	29,74	29,74	0,52
BAIRRO.BARRA DO CEARA	29,67	29,67	0,52
SITUACAO.LOTE.IPTU.Encravado	29,62	29,62	0,52
BAIRRO.VILA VELHA	29,62	29,62	0,52
BAIRRO.PARANGABA	29,62	29,62	0,52
BAIRRO.MOURA BRASIL	29,57	29,57	0,52
BAIRRO.FATIMA	29,5	29,5	0,52
BAIRRO.OLAVO OLIVEIRA	29,47	29,47	0,52
BAIRRO.ALVARO WEYNE	29,43	29,43	0,52
BAIRRO.PARREAO	29,28	29,28	0,51
BAIRRO.AEROPORTO	29,18	29,18	0,51
PADRAO.EDIFICACAO.Alto nivel 3	29,11	29,11	0,51
BAIRRO.BARROSO	29,11	29,11	0,51

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
PADRAO.EDIFICACAO.Normal 1	29,01	29,01	0,51
PADRAO.EDIFICACAO.Luxo 1	28,92	28,92	0,51
BAIRRO.PRAIA DE IRACEMA	28,91	28,91	0,51
BAIRRO.JOQUEI CLUBE	28,9	28,9	0,51
SITUACAO.LOTE.IPTU.Vila	28,88	28,88	0,51
OCUPACAO.SEM	28,83	28,83	0,51
BAIRRO.MESSEJANA	28,77	28,77	0,5
BAIRRO.PRAIA DO FUTURO II	28,76	28,76	0,5
USO.ESPECIFICO.Instrucao	28,76	28,76	0,5
BAIRRO.MANOEL DIAS BRANCO	28,74	28,74	0,5
BAIRRO.CONJUNTO CEARA I	28,73	28,73	0,5
BAIRRO.PRAIA DO FUTURO I	28,72	28,72	0,5
BAIRRO.PARQUE SAO JOSE	28,61	28,61	0,5
BAIRRO.CARLITO PAMPLONA	28,6	28,6	0,5
BAIRRO.CRISTO REDENTOR	28,6	28,6	0,5
BAIRRO.NOVO MONDUBIM	28,59	28,59	0,5
BAIRRO.FLORESTA	28,58	28,58	0,5

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
PADRAO.EDIFICACAO.Luxo 2	28,56	28,56	0,5
BAIRRO.JARDIM IRACEMA	28,48	28,48	0,5
BAIRRO.CANINDEZINHO	28,4	28,4	0,5
SITUACAO.LOTE.IPTU.	28,38	28,38	0,5
BAIRRO.JARDIM GUANABARA	28,33	28,33	0,5
TIPO.LOGRADOURO.ESTRADA	28,3	28,3	0,5
BAIRRO.DE LOURDES	28,29	28,29	0,5
BAIRRO.CAIS DO PORTO	28,29	28,29	0,5
TIPO.LOGRADOURO.VILA	28,18	28,18	0,49
TIPO.LOGRADOURO.PRACA	28,17	28,17	0,49
BAIRRO.AUTRAN NUNES	28,13	28,13	0,49
TIPO.LOGRADOURO.RODOVIA	27,99	27,99	0,49
BAIRRO.CONJUNTO CEARA II	27,99	27,99	0,49
SITUACAO.LOTE.IPTU.Gleba	27,98	27,98	0,49
BAIRRO.CONJUNTO ESPERANCA	27,97	27,97	0,49
PADRAO.EDIFICACAO.Baixo nivel 3	27,96	27,96	0,49
BAIRRO.CIDADE DOS FUNCIONARIOS	27,91	27,91	0,49
BAIRRO.DEMOCRITO ROCHA	27,87	27,87	0,49
BAIRRO.RODOLFO TEOFILO	27,79	27,79	0,49

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
USO.ESPECIFICO.Prestacao	27,68	27,68	0,49
USO.ESPECIFICO.Do lazer	27,65	27,65	0,49
BAIRRO.SABIAGUABA	27,64	27,64	0,49
BAIRRO.SIQUEIRA	27,63	27,63	0,48
BAIRRO.COUTO FERNANDES	27,6	27,6	0,48
BAIRRO.PARQUE SANTA ROSA	27,58	27,58	0,48
BAIRRO.SAO BENTO	27,48	27,48	0,48
BAIRRO.CONJUNTO PALMEIRAS	27,47	27,47	0,48
BAIRRO.	27,47	27,47	0,48
USO.ESPECIFICO.Institucional	27,45	27,45	0,48
TIPO.LOGRADOURO.TRAVESSA	27,39	27,39	0,48
TIPO.LOGRADOURO.BECO	27,22	27,22	0,48
PADRAO.EDIFICACAO.Baixo nivel 1	27,19	27,19	0,48
USO.ESPECIFICO.Saude	27,17	27,17	0,48
BAIRRO.CAJAZEIRAS	27,16	27,16	0,48
PADRAO.EDIFICACAO.Baixo nivel 2	27,12	27,12	0,48
OCUPACAO.Ruinas/demolicao	27,07	27,07	0,48
BAIRRO.CURIO	27,04	27,04	0,47
USO.ESPECIFICO.Fechado	27,02	27,02	0,47
BAIRRO.MANOEL SATIRO	27,01	27,01	0,47
SITUACAO.LOTE.IPTU.Quadra	27	27	0,47
USO.ESPECIFICO.Religioso	26,95	26,95	0,47
BAIRRO.FARIAS BRITO	26,87	26,87	0,47

Tabela 9 – Importância de variáveis para o modelo RNA

(Continuação)

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
BAIRRO.GENIBAU	26,83	26,83	0,47
BAIRRO.PARQUE PRESIDENTE VARGAS	26,8	26,8	0,47
BAIRRO.COACU	26,73	26,73	0,47
BAIRRO.PAN AMERICANO	26,72	26,72	0,47
BAIRRO.PAUPINA	26,59	26,59	0,47
NUMERO.FRENTES	26,41	26,41	0,46
USO.ESPECIFICO.Hotelaria	26,39	26,39	0,46
USO.ESPECIFICO.Comunicacao	26,13	26,13	0,46
USO.ESPECIFICO.Sem	26,02	26,02	0,46
BAIRRO.VILA PERY	26	26	0,46
BAIRRO.GRANJA PORTUGAL	25,87	25,87	0,45
BAIRRO.VILA ELLERY	25,68	25,68	0,45
USO.ESPECIFICO.Industrial	25,59	25,59	0,45
BAIRRO.PREFEITO JOSE VALTER	25,11	25,11	0,44
BAIRRO.missing(NA)	0	0	0
USO.ESPECIFICO.missing(NA)	0	0	0
PADRAO.EDIFICACAO.missing(NA)	0	0	0
TIPO.LOGRADOURO.missing(NA)	0	0	0
SITUACAO.LOTE.IPTU.missing(NA)	0	0	0
OCUPACAO.missing(NA)	0	0	0

Fonte o Autor, 2019.

APÊNDICE E – *IMPORTÂNCIA DE VARIÁVEIS PARA O MODELO RF*

Variable	*Relative Importance*	*Scaled Importance*(%)	*Percentage*(%)
VALOR.VENAL.IPTU.NO.EXERCICIO	757735,94	100	29,78
AREA.EDIFICADA.GI	641068,19	84,6	25,19
BAIRRO	433643,97	57,23	17,04
AREA.TERRENO.GI	160905,5	21,24	6,32
DATA.CADASTRO.GI	108543,34	14,32	4,27
FATOR.EDIFICACAO	103181,11	13,62	4,05
PADRAO.EDIFICACAO	73784,16	9,74	2,9
X	54153,25	7,15	2,13
TESTADA.PRINCIPAL	51194,54	6,76	2,01
DATA.CONSTRUCAO	38968,8	5,14	1,53
Y	34079,62	4,5	1,34
FATOR.LOTE	23320,43	3,08	0,92
FRACAO.IDEAL	20141,7	2,66	0,79
NUMERO.PAVIMENTOS	12023,68	1,59	0,47
NUM.UNIDADES.LOTE	8715,04	1,15	0,34
TIPO.LOGRADOURO	6703,56	0,88	0,26
SITUACAO.LOTE.IPTU	6129,95	0,81	0,24
USO.ESPECIFICO	5367,12	0,71	0,21
NUMERO.FRENTES	3983,17	0,53	0,16
OCUPACAO	1086,07	0,14	0,04
AREA.PRESERVACAO.GI	80,31	0,01	0

Fonte: o Autor, 2019.

APÊNDICE F – IMPORTÂNCIA DE VARIÁVEIS PARA O MODELO RLM

variable	*Estimate*	*Std.*	*Error*	*value*
(Intercept)	-616100	11090	-55564	2,00E-16
DATA.CADASTRO.GI	143,9	2,241	64191	2,00E-16
DATA.CONSTRUCAO	5,084	0,2781	18280	2,00E-16
NUMERO.PAVIMENTOS	-15,12	5,762	-2624	0.00870
NUMERO.FRENTES	23,15	5,693	4067	4.79e-05
NUM.UNIDADES.LOTE	-3,988	0,31	-12861	2,00E-16
TESTADA.PRINCIPAL	3,481	0,2371	14684	2,00E-16
FATOR.EDIFICACAO	1718	42,92	40021	2,00E-16
FATOR.LOTE	-963,9	60,2	-16011	2,00E-16
AREA.TERRENO.GI	-0,004458	0,002236	-1994	0.04618
AREA.PRESERVACAO.GI	0,5516	0,2077	2656	0.00791
FRACAO.IDEAL	20,67	13,33	1551	0.12100
AREA.EDIFICADA.GI	4,899	0,04871	100592	2,00E-16
VALOR.VENAL.IPTU.NO.EXERCICIO	0,00001207	0,00001201	1005	0.31509
X	0,03658	0,0007231	50592	2,00E-16
Y	0,03203	0,001003	31948	2,00E-16

Fonte: o Autor, 2019.

APÊNDICE G – *CÓDIGOS BASE PARA O ALGORITMO PREDITOR*

Código-fonte 1 – Bibliotecas importadas e variaveis base

```
###### Iniciais ##############
rm(list=ls()) #Limpa memoria do console R
library(data.table) #Leitor e escritor de arquivos
require(dplyr) #processador de funcoes de filtro
library(h2o) #Modelos aprendizagem de maquina
library(lubridate) #Conversor de datas
library(ggplot2) #Biblioteca para graficos
library(MLmetrics) # Calculo do erro medio percentual
library(tidyverse) # Modelo Linear
library(ggpubr)
library(hrbrthemes)
library(viridis)
semente = 100000 #Semente
set.seed(semente) # Colocando Amostra fixa para garantir
   reprodutibilidade dos resultados
h2o.init(nthreads=-1) ## Inicia Cluster h2o com todos os
   nucleos disponiveis
num_folds = 10
test_name = "Dados_Set_Final_Definitive" #Nomeia amostra
```

Código-fonte 2 – Tratamento de dados brutos

```
#Conversao de dados brutos, primeiramente desconsiderou-se
   da base os dados pertinentes ao A PMF e os diretamente
   dependentes da Lei de uso e ocupacao, como: Quadra, lote
   , distrito, situacao GI, e TIPO_GI.
#Em seguida as variaveis categoricas foram convertidas para
    fatores, as numericas para numeros e das datas foram
   retiradas apenas os anos.
```

```
3 #Para a variavel a ser prevista "VALOR_BASE_CALCULO_ITBI"
    foi adotada uma escala logaritmica. Por ultimo, foi
    feito o filtro apenas para unidades residenciais do tipo
     "Casa".
4 #filtrados os dados para apenas os apos 2011 e desprezando
    entradas repetidas no banco de dados.
5
6 df_base_geral <- fread(paste0(dirname(rstudioapi::
    getSourceEditorContext()$path),"/Dados Base/ITBI_2009_
    2016.csv"), sep=";", dec=",") %>% select(-c(TIPO_GI,
    EXERCICIO_GI,SITUACAO_GI,DISTRITO,QUADRA,LOTE)) %>%
    filter(CLASSIF_ARQUITETONICA == "Casa") %>% distinct()
    %>% mutate(X = as.numeric(X), Y = as.numeric(Y), FATOR_
    EDIFICACAO = as.numeric(FATOR_EDIFICACAO), FATOR_LOTE =
    as.numeric(FATOR_LOTE), AREA_PRESERVACAO_GI = as.numeric
    (AREA_PRESERVACAO_GI), FRACAO_IDEAL = as.numeric(FRACAO_
    IDEAL), VALOR_VENAL_IPTU_NO_EXERCICIO = as.numeric(VALOR
    _VENAL_IPTU_NO_EXERCICIO), VALOR_BASE_CALCULO_ITBI = as.
    numeric(VALOR_BASE_CALCULO_ITBI),DATA_CONSTRUCAO = year(
    dmy(DATA_CONSTRUCAO)),DATA_CADASTRO_GI = year(dmy_hm(
    DATA_CADASTRO_GI)),TIPO_LOGRADOURO = as.factor(TIPO_
    LOGRADOURO),BAIRRO = as.factor(BAIRRO),TIPO_IMOVEL = as.
    factor(TIPO_IMOVEL),USO_ESPECIFICO = as.factor(USO_
    ESPECIFICO),OCUPACAO = as.factor(OCUPACAO),NUMERO_
    PAVIMENTOS = as.factor(NUMERO_PAVIMENTOS),NUMERO_FRENTES
     = as.factor(NUMERO_FRENTES),NUM_UNIDADES_LOTE = as.
    factor(NUM_UNIDADES_LOTE),SITUACAO_LOTE_IPTU = as.factor
    (SITUACAO_LOTE_IPTU),PADRAO_EDIFICACAO = as.factor(
    PADRAO_EDIFICACAO)) %>% filter(DATA_CADASTRO_GI >= 2011)
     %>% select(-c(CLASSIF_ARQUITETONICA, TIPO_IMOVEL))
7
8 histograma_sem_tratamento <- ggplot(df_base_geral, aes(
```

```
   VALOR_BASE_CALCULO_ITBI)) + geom_histogram(color="
   #122625", fill = "#8d6a9f",alpha=0.6, position = '
   identity', bins = 30) + theme_ipsum() + xlab("Valor ITBI
    (R$)") + ylab("Frequencia n") + ggtitle("Histograma de
   valores ITBI\nSem Tratamento") + theme(plot.title =
   element_text(hjust = 0.5)) + theme(axis.title.x =
   element_text(hjust = 0.5),axis.title.y = element_text(
   hjust = 0.5))
histograma_sem_tratamento
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/histograma_sem_
   tratamento.png"))
dev.off ()

df_base_geral <- df_base_geral%>% mutate(VALOR_BASE_CALCULO
   _ITBI = log(VALOR_BASE_CALCULO_ITBI))
#Embaralhamento de dados do dataframe
df_base_geral <- df_base_geral[sample(nrow(df_base_geral))
   ,]

# Plot dados de comparacao de area com outliers
db_com_outiers <- ggplot(df_base_geral, aes(x=AREA_
   EDIFICADA_GI, y=VALOR_BASE_CALCULO_ITBI)) + geom_point(
   aes(color= VALOR_BASE_CALCULO_ITBI)) + geom_smooth(
   method=lm , color="blue", linetype = "twodash", size =
   1.5) + theme_ipsum() + ylim(min(df_base_geral$VALOR_BASE
   _CALCULO_ITBI),max(df_base_geral$VALOR_BASE_CALCULO_ITBI
   ))  + theme(axis.title.x = element_text(hjust = 0.5),
   axis.title.y = element_text(hjust = 0.5)) + xlim(0,1000)
    + ggtitle("Valor de ITBI por area\nCom Outliers") +
   xlab ("Area edificada (m2)") + ylab("Valor ITBI (R$) -
   Escala Logaritmica")  + theme(plot.title = element_text(
```

```
   hjust = 0.5))  + scale_color_viridis(option = "inferno")
    + theme(legend.position = "bottom")
db_com_outiers

dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/db_com_outliers
   .png"))
dev.off ()
#boxplot sem
#boxplot com outliers
box_plot_com_outliers <- ggplot(df_base_geral, aes(x="" ,y=
   VALOR_BASE_CALCULO_ITBI)) + geom_boxplot(varwidth = TRUE
   ,color="#122625", fill = "#3a606e",alpha=0.6) + ggtitle(
   "BoxPlot Valores ITBI\nCom Outliers") + xlab ("
   Transacoes") + ylab("Valor ITBI (R$) - Escala
   Logaritmica")  + theme(plot.title = element_text(hjust =
    0.5))
box_plot_com_outliers
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/boxplot_com_
   outliers.png"))
dev.off ()
#boxplot sem

# histograma com outliers
histograma <- ggplot(df_base_geral, aes(VALOR_BASE_CALCULO_
   ITBI)) + geom_histogram(color="#122625", fill = "#3a606e
   ",alpha=0.6, position = 'identity', bins = 30) + theme_
   ipsum() + xlab("Valor ITBI (R$) - Escala Logaritmica") +
    ylab("Frequencia n") + ggtitle("Histograma de valores
   ITBI\nCom Outliers") + theme(plot.title = element_text(
   hjust = 0.5)) + theme(axis.title.x = element_text(hjust
```

```
   = 0.5),axis.title.y = element_text(hjust = 0.5))
histograma
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/histograma_com_
   outliers.png"))
dev.off ()
#boxplot sem

#Retirada dos outliers da amostra
df_base_geral <- df_base_geral[!df_base_geral$VALOR_BASE_
   CALCULO_ITBI %in% as.data.frame(matrix(boxplot(df_base_
   geral$VALOR_BASE_CALCULO_ITBI, plot=FALSE)$out))$V1,]
```

Código-fonte 3 – Plotagem de dados

```
#Plot of data sem outliers.

# Plot dados de comparacao de area com outliers
db_sem_outiers <- ggplot(df_base_geral, aes(x=AREA_
   EDIFICADA_GI, y=VALOR_BASE_CALCULO_ITBI)) + geom_point(
   aes(color= VALOR_BASE_CALCULO_ITBI)) + geom_smooth(
   method=lm , color="blue", linetype = "twodash", size =
   1.5) + theme_ipsum() + ylim(min(df_base_geral$VALOR_BASE
   _CALCULO_ITBI),max(df_base_geral$VALOR_BASE_CALCULO_ITBI
   )) + theme(axis.title.x = element_text(hjust = 0.5),
   axis.title.y = element_text(hjust = 0.5)) + xlim(0,1000)
    + ggtitle("Valor de ITBI por area\nSem Outliers") +
   xlab ("Area edificada (m2)") + ylab("Valor ITBI (R$) -
   Escala Logaritmica") + theme(plot.title = element_text(
   hjust = 0.5)) + scale_color_viridis(option = "inferno")
    + theme(legend.position = "bottom")
db_sem_outiers
```

```
6
7  dev.copy(png,filename= paste0(dirname(rstudioapi::
       getSourceEditorContext()$path),"/Figuras/db_sem_outliers
       .png"))
8  dev.off ()
9  #boxplot sem outliers
10 box_plot_sem_outliers <- ggplot(df_base_geral, aes(x="" ,y=
       VALOR_BASE_CALCULO_ITBI)) + geom_boxplot(varwidth = TRUE
       , color="#69b3a2", fill = "#e0fbfc",alpha=0.6) + ggtitle
       ("BoxPlot Valores ITBI\nSem Outliers") + xlab ("
       Transacoes") + ylab("Valor ITBI (R$) - Escala
       Logaritmica")  + theme(plot.title = element_text(hjust =
        0.5))
11 box_plot_sem_outliers
12 dev.copy(png,filename= paste0(dirname(rstudioapi::
       getSourceEditorContext()$path),"/Figuras/boxplot_sem_
       outliers.png"))
13 dev.off ()
14
15 #Histograma sem outiers
16 histograma_sem_outliers <- ggplot(df_base_geral, aes(VALOR_
       BASE_CALCULO_ITBI)) + geom_histogram(color="#69b3a2",
       fill = "#e0fbfc",alpha=0.6, position = 'identity', bins
       = 30) + xlab("Valor ITBI (R$) - Escala Logaritmica") +
       theme_ipsum()  + ylab("Frequencia n") + ggtitle("
       Histograma de valores ITBI\nSem Outliers") + theme(plot.
       title = element_text(hjust = 0.5)) + theme(axis.title.x
       = element_text(hjust = 0.5),axis.title.y = element_text(
       hjust = 0.5))
17 histograma_sem_outliers
18 dev.copy(png,filename= paste0(dirname(rstudioapi::
       getSourceEditorContext()$path),"/Figuras/histograma_sem_
```

```
   outliers.png"))
dev.off ()


ggarrange(db_com_outiers,db_sem_outiers)
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/dados.png"),
   width = 1280, height = 720)
dev.off ()

ggarrange(histograma,histograma_sem_outliers)
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/histogramas.png
   "), width = 1280, height = 720)
dev.off ()

ggarrange(box_plot_com_outliers,box_plot_sem_outliers)
dev.copy(png,filename= paste0(dirname(rstudioapi::
   getSourceEditorContext()$path),"/Figuras/boxplots.png"),
    width = 1280, height = 720)
dev.off ()
```

Código-fonte 4 – Codigo base para Regressão Linear Múltipla (RLM)

```
### BASE RLM ###
rlm_train_base <- df_base_geral[sample(1:nrow(df_base_geral
   ), size = round(0.70 * nrow(df_base_geral),0) ),] #
   pegando 70% das observacoes para a amostra
rlm_test_base <- anti_join(df_base_geral,rlm_train_base)

rlm_train_base <- rlm_train_base %>% select(-c(TIPO_
   LOGRADOURO,BAIRRO,USO_ESPECIFICO,OCUPACAO,SITUACAO_LOTE_
```

```
   IPTU,PADRAO_EDIFICACAO)) %>% mutate(NUMERO_PAVIMENTOS =
   as.numeric(NUMERO_PAVIMENTOS), NUMERO_FRENTES = as.
   numeric(NUMERO_FRENTES),NUM_UNIDADES_LOTE = as.numeric(
   NUM_UNIDADES_LOTE))
for(i in 1:ncol(rlm_train_base)) {
  rlm_train_base[ , i][is.na(rlm_train_base[ , i])] <- mean
     (rlm_train_base[ , i], na.rm = TRUE)
}

rlm_test_base <- rlm_test_base %>% select(-c(TIPO_
   LOGRADOURO,BAIRRO,USO_ESPECIFICO,OCUPACAO,SITUACAO_LOTE_
   IPTU,PADRAO_EDIFICACAO)) %>% mutate(NUMERO_PAVIMENTOS =
   as.numeric(NUMERO_PAVIMENTOS), NUMERO_FRENTES = as.
   numeric(NUMERO_FRENTES),NUM_UNIDADES_LOTE = as.numeric(
   NUM_UNIDADES_LOTE))
for(i in 1:ncol(rlm_test_base)) {
  rlm_test_base[ , i][is.na(rlm_test_base[ , i])] <- mean(
     rlm_test_base[ , i], na.rm = TRUE)
}

#PCriacao do modelo de regressao linear multipla
rlm_model <- lm(VALOR_BASE_CALCULO_ITBI  ~.,  data = rlm_
   train_base)
#plot(rlm_model)
#Afericao do erro medio percentual do modelo
rlm_mape <- MAPE(as.matrix(predict(rlm_model, rlm_test_base
   )), rlm_test_base$VALOR_BASE_CALCULO_ITBI)

#Escrita de Performance em termos de MAE, MSE, RMSE e RMSLE
    para o modelo
sink(paste0(dirname(rstudioapi::getSourceEditorContext()$
   path),"/Dados Base/",test_name,"_rlm.txt"))
```

```
23
24 cat(paste("MLMETRICS: Regressao Linear Multipla","",paste("
     MSE:  ", round(MSE(as.matrix(predict(rlm_model, rlm_test
     _base)), rlm_test_base$VALOR_BASE_CALCULO_ITBI),8), sep
     = "  "),paste("RMSE:  ", round(RMSE(as.matrix(predict(
     rlm_model, rlm_test_base)), rlm_test_base$VALOR_BASE_
     CALCULO_ITBI),8), sep = "  "), paste("MAE:  ", round(MAE
     (as.matrix(predict(rlm_model, rlm_test_base)), rlm_test_
     base$VALOR_BASE_CALCULO_ITBI),8), sep = "  "), paste("
     RMSLE:  ", round(RMSLE(as.matrix(predict(rlm_model, rlm_
     test_base)), rlm_test_base$VALOR_BASE_CALCULO_ITBI),8),
     sep = "  "), sep = '\n')[1])
25 summary(rlm_model)
26 sink(NULL)
```

Código-fonte 5 – Código base para Regressão Gaussiana (GLM)

```
1 glm_train_base <- df_base_geral[sample(1:nrow(df_base_geral
    ), size = round(0.70 * nrow(df_base_geral),0) ),] #
    pegando 70% das observacoes para a amostra
2 glm_test_base <- anti_join(df_base_geral,glm_train_base) #
    Usando o resto da amostra para teste
3
4 #Importando arquivos para cluster h2o
5 fwrite(glm_train_base, "glm_train_base.csv")
6 fwrite(glm_test_base,"glm_test_base.csv")
7 glm_train_base <- h2o.importFile("glm_train_base.csv")
8 glm_test_base <- h2o.importFile("glm_test_base.csv")
9
10 #Premissas para modelo de regressao Gaussiana:
11 #Distribuicao Gaussiana, metodo = Iteratively Reweighted
    Least Squares Method
```

```
general_linear_model <- h2o.glm(model_id = "glm_first_model",training_frame = glm_train_base, validation_frame = glm_test_base, y= "VALOR_BASE_CALCULO_ITBI", seed = semente,  nfolds = num_folds, family = "gaussian",solver = "IRLSM")

#Escrita de importancia das variaveis para o modelo
fwrite(as.data.frame(h2o.varimp(general_linear_model)), paste0(dirname(rstudioapi::getSourceEditorContext()$path),"/Dados Base/",test_name,"_glm_variables.csv"), sep = ";", dec=",")

#Escrita de Performance em termos de MAE, MSE, RMSE, RMSLE e R2 para o modelo
h2o.performance(general_linear_model, newdata = glm_test_base)
sink(paste0(dirname(rstudioapi::getSourceEditorContext()$path),"/Dados Base/",test_name,"_glm.txt"))
h2o.performance(general_linear_model, newdata = glm_test_base)
sink(NULL)

#Afericao do erro medio percentual do modelo
glm_mape <- MAPE(as.matrix(predict(general_linear_model, glm_test_base)), as.matrix(fread("glm_test_base.csv",sep = ",",dec= ".")$VALOR_BASE_CALCULO_ITBI))
```

Código-fonte 6 – Codigo base para *Random Forests* (RF)

```
###### BASE RF ######
rf_train_base <- df_base_geral[sample(1:nrow(df_base_geral), size = round(0.70 * nrow(df_base_geral),0) ),] #
```

```
   pegando 70% das observacoes para a amostra
rf_test_base <- anti_join(df_base_geral,rf_train_base)#
   Usando o resto da amostra para teste

#Importando arquivos para cluster h2o
fwrite(rf_train_base, "rf_train_base.csv")
fwrite(rf_test_base,"rf_test_base.csv")
rf_train_base <- h2o.importFile("rf_train_base.csv")
rf_test_base <- h2o.importFile("rf_test_base.csv")

#Premissas para modelo de Random Forests:
#Histograma uniforme e 300 arvores criadas, e usando
   adaptacao uniforme para o tipo do histograma.
random_forest_model<- h2o.randomForest(model_id = "rf_first
   _model",training_frame = rf_train_base, validation_frame
    = rf_test_base, y= "VALOR_BASE_CALCULO_ITBI", seed =
   semente, nfolds = num_folds, ntrees = 300, histogram_
   type = "UniformAdaptive")
plot(random_forest_model)
#Escrita de importancia das variaveis para o modelo
fwrite(as.data.frame(h2o.varimp(random_forest_model)),
   paste0(dirname(rstudioapi::getSourceEditorContext()$path
   ),"/Dados Base/",test_name,"_rf_variables.csv"), sep = "
   ;", dec=",")

#Escrita de Performance em termos de MAE, MSE, RMSE e RMSLE
    para o modelo
h2o.performance(random_forest_model, newdata = rf_test_base
   )
sink(paste0(dirname(rstudioapi::getSourceEditorContext()$
   path),"/Dados Base/",test_name,"_rf.txt"))
h2o.performance(random_forest_model, newdata = rf_test_base
```

```
   )
sink(NULL)

#Afericao do erro medio percentual do modelo
rf_mape <- MAPE(as.matrix(predict(random_forest_model, rf_
   test_base)), as.matrix(fread("rf_test_base.csv",sep = ",
   ",dec= ".")$VALOR_BASE_CALCULO_ITBI))
```

Código-fonte 7 – Codigo base para *Gradient Boosting Machine* (GBM)

```
### BASE GBM ###
gbm_train_base <- df_base_geral[sample(1:nrow(df_base_geral
   ), size = round(0.70 * nrow(df_base_geral),0) ),] #
   pegando 70% das observacoes para a amostra
gbm_test_base <- anti_join(df_base_geral,gbm_train_base)

fwrite(gbm_train_base, "gbm_train_base.csv")
fwrite(gbm_test_base,"gbm_test_base.csv")
gbm_train_base <- h2o.importFile("gbm_train_base.csv")
gbm_test_base <- h2o.importFile("gbm_test_base.csv")

#Premissas para modelo de Gradient boosting machine:
#Distribuicao Gaussiana, taxa de aprendizagem = 0,2, e 300
   arvores criadas.
gradient_boost_model <- h2o.gbm(model_id = "gbm_first_model
   ",training_frame = gbm_train_base, validation_frame =
   gbm_test_base, y= "VALOR_BASE_CALCULO_ITBI", seed =
   semente, learn_rate = 0.2, nfolds = num_folds, ntrees =
   300 ,distribution = "gaussian")
plot(gradient_boost_model)

#Escrita de importancia das variaveis para o modelo
```

```
fwrite(as.data.frame(h2o.varimp(gradient_boost_model)),
   paste0(dirname(rstudioapi::getSourceEditorContext()$path
   ),"/Dados Base/",test_name,"_gbm_variables.csv"), sep =
   ";", dec=",")

#Escrita de Performance em termos de MAE, MSE, RMSE e RMSLE
    para o modelo
h2o.performance(gradient_boost_model, newdata = gbm_test_
   base)
sink(paste0(dirname(rstudioapi::getSourceEditorContext()$
   path),"/Dados Base/",test_name,"_gbm.txt"))
h2o.performance(gradient_boost_model, newdata = gbm_test_
   base)
sink(NULL)

#Afericao do erro medio percentual do modelo
gbm_mape <- MAPE(as.matrix(predict(gradient_boost_model,
   gbm_test_base)), as.matrix(fread("gbm_test_base.csv",sep
    = ",",dec= ".")$VALOR_BASE_CALCULO_ITBI))
```

Código-fonte 8 – Código base para Redes Neurais Artificiais (RNA)

```
###BASE RNA ###
rna_train_base <- df_base_geral[sample(1:nrow(df_base_geral
   ), size = round(0.70 * nrow(df_base_geral),0) ),] #
   pegando 70% das observacoes para a amostra
rna_test_base <- anti_join(df_base_geral,rna_train_base) #
   Usando o resto da amostra para teste

#Importando arquivos para cluster h2o
fwrite(rna_train_base, "rna_train_base.csv")
fwrite(rna_test_base,"rna_test_base.csv")
```

```
8  rna_train_base <- h2o.importFile("rna_train_base.csv")
9  rna_test_base <- h2o.importFile("rna_test_base.csv")
10
11
12 #Premissas para modelo de redes neurais artificiais:
13 #Distribuicao Gaussiana, funcao perda = funcao quadratica,
     numero de epocas = 10
14 rede_neural <- h2o.deeplearning( model_id="dl_model_first",
     y= "VALOR_BASE_CALCULO_ITBI", training_frame = rna_train
     _base, validation_frame = rna_test_base, nfolds = num_
     folds, ignore_const_cols = TRUE, seed = semente,
     standardize = TRUE, epochs = 10, distribution = "
     gaussian", loss= "Quadratic", missing_values_handling =
     "MeanImputation")
15
16 #Escrita de importancia das variaveis para o modelo
17 fwrite(as.data.frame(h2o.varimp(rede_neural)),paste0(
     dirname(rstudioapi::getSourceEditorContext()$path),"/
     Dados Base/",test_name,"_rna_variables.csv"), sep = ";",
      dec=",")
18
19 #Escrita de Performance em termos de MAE, MSE, RMSE e RMSLE
      para o modelo
20 h2o.performance(rede_neural, newdata = rna_test_base)
21 sink(paste0(dirname(rstudioapi::getSourceEditorContext()$
     path),"/Dados Base/",test_name,"_rna.txt"))
22 h2o.performance(rede_neural, newdata = rna_test_base)
23 sink(NULL)
24
25 #Afericao do erro medio percentual do modelo
26 rna_mape <- MAPE(as.matrix(predict(rede_neural, rna_test_
     base)), as.matrix(fread("rna_test_base.csv",sep = ",",
```

```
  dec= ".")$VALOR_BASE_CALCULO_ITBI))
```