UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Precificação de Imóveis com Machine Learning

Sérgio Ricardo Ribeiro Alencar

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Precificação de Imóveis com Machine Learning

Sérgio Ricardo Ribeiro Alencar

USP - São Carlos

2022

Sérgio Ricardo Ribeiro Alencar

Precificação de Imóveis com Machine Learning

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Machine Learning

Orientadora: Prof. Dr. Glauco Caurin

USP - São Carlos

2022

R368p

Ribeiro Alencar, Sérgio Ricardo
Precificação de Imóveis com Machine Learning / Sérgio Ricardo Ribeiro Alencar; orientador Glauco Caurin. -- São Carlos, 2022.
40 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Machine Learning. 2. Precificação de Imóveis. 3. Mercado Imobiliário. I. Caurin, Glauco, orient. II. Título.

DEDICATÓRIA

Agradeço primeiramente ao meu orientador, o Professor Glauco Caurin por ter aceitado acompanhar-me neste projeto. Agradeço também à minha família pela compreensão, carinho e apoio incansável.

EPÍGRAFE

"The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning…. Data-driven predictions can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves."

Silver (2012)

RESUMO

ALENCAR, R. R. S. **Precificação de Imóveis com Machine Learning**. 2022. 40 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Imóveis são uma das alternativas de investimentos mais populares entre os brasileiros. Dos brasileiros com renda superior a R$ 10 mil, 57% ainda não possuem imóveis como investimento, mas desejam fazer uma aquisição com esse objetivo. Porém, é preciso cautela na hora da decisão de compra, já que o principal fator para ter um bom retorno realizando este tipo de investimento é comprar o imóvel por um valor justo ou abaixo do mercado, ou comprar imóveis em áreas com alto potencial de valorização. Da mesma forma, é preciso conhecer bem o imóvel e a região para precificá-lo corretamente na hora da venda. Com o objetivo de servir ambos os públicos foi desenvolvida uma ferramenta que por meio da coleta e da estruturação de dados e com posterior aplicação de machine learning é capaz de precificar o valor de um imóvel. Neste produto, o machine learning foi usado para predizer valores baseado em imóveis similares provenientes da base de dados de Melbourne, na Austrália, mas este modelo pode ser aplicado em qualquer outro banco de dados. Após testes dos modelos *Random Forest*, *KNN* e regressão linear foi demonstrado mediante a análise de resultados que o modelo com resultado mais preciso de predição foi o *Random Forest* predizendo os preços dos imóveis com erro de 13,34%.

Palavras-chave: Machine Learning 1; Precificação de Imóveis 2; Mercado Imobiliário 3.

ABSTRACT

ALENCAR, R. R. S. **Precificação de Imóveis com Machine Learning**. 2022. 40 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Real estate is one of the most popular investment alternatives in Brazil. Of the Brazilians with income higher than R$ 10,000, 57% still do not have properties, but want to acquire a property with this purpose. However, caution is needed when making the purchase decision, since the main factor to have a good return on this type of investment is to buy the property at a fair or below market value, or to buy properties in areas with high potential for recovery. Likewise, it is necessary to know the properties and the region well to correctly know the fair sales price. able to price the value of a property. To serve both audiences, a tool was developed that, through the collection and structuring of data and with subsequent application of machine learning, can price the value of a property. In this product, machine learning was used to predict values based on similar properties from Melbourne, Australia database. Nevertheless, this model can be applied to any other database. After testing the Random Forest, KNN and linear regression models, it was demonstrated through the analysis of results that the model with the most accurate prediction result was the Random Forest predicting property prices with an error of 13.34%.

Keywords: Machine Learning 1; Real Estate Pricing 2; Real Estate 3.

SUMÁRIO

1 INTRODUÇÃO

A título de investimento, os imóveis são uma das alternativas mais populares entre os brasileiros. De acordo com a Brain Inteligência Estratégica (CBIC, 2021), empresa de pesquisa e consultoria em negócios, em sua pesquisa divulgada em março de 2021 com 6 mil brasileiros com renda superior a R$ 10 mil, 57% daqueles que ainda não possuem imóveis como investimento, mas desejam fazer uma aquisição com esse objetivo. Destes, 54% pretende obter rentabilidade por meio de aluguéis, enquanto 32% pretende tê-los como reserva de valor e outros 14% querem comprar para revender.

Porém, para todos os fins citados, é preciso cautela na hora da decisão de compra, já que o principal fator para ter um bom retorno realizando este tipo de investimento é comprar o imóvel por um valor justo ou abaixo do mercado, ou comprar imóveis em áreas com alto potencial de valorização, seja por ser uma região com melhorias de infraestrutura, seja por ser uma região de expansão natural de alguma cidade.

Portanto, para otimizar a tomada de decisão de investir em um imóvel é necessário possuir informações sobre os valores praticados em uma determinada região e comparar o imóvel em questão com este valor. Entretanto, hoje no mercado são limitadas as opções rápidas e gratuitas para se avaliar um imóvel desta forma, ou para compará-lo com outros imóveis similares. Para realizar tais análises seria necessário bastante tempo e dedicação para conseguir um número razoável de propriedades comparáveis, ou a ajuda de um profissional com bastante tempo de atuação em determinada região, como corretores e como imobiliárias experientes, cenário no qual pode haver enviesamento de informações e conflito de interesse. Assim, torna-se um processo bastante moroso o de buscar por oportunidades no mercado imobiliário na maioria das localidades do país, o que dificulta a tomada de decisão dos investidores e reduz a quantidade e a velocidade de possíveis negócios que poderiam acontecer com mais frequência.

1.1 A inovação

Com base na grande relevância que o mercado imobiliário tem para o Brasil e na cultura brasileira de investir nestes bens pensando em gerar renda e em criar patrimônio, faz-se necessário para estas pessoas o uso de ferramentas que lhes possibilitem saber qual o preço justo de um imóvel à venda como também por quanto vender seus imóveis quando fizer sentido.

Desta forma, seria desenvolvida uma plataforma que, por meio da coleta e da estruturação de dados com posterior aplicação de *machine learning*, conseguiria precificar o valor de um imóvel em qualquer cidade do país baseada em dados dinâmicos de diversas plataformas distintas de anúncio de imóveis. Neste produto, o machine learning seria usado para predizer valores baseado em imóveis similares. Hoje já há plataformas que precificam imóveis em grandes capitais do país, mas a inovação viria da capacidade dinâmica de replicar o modelo também para cidades de médio e de pequeno porte e, além disso, esta plataforma poderia servir para monetizar os *leads* adquiridos por meio desta ferramenta para outras plataformas especializadas em compra e venda de imóveis, propósito para o qual ainda não foram encontrados concorrentes. Por último, este modelo também poderá servir de base para uma outra inovação de mercado, que seria um avaliador de preços capaz de determinar se um imóvel está acima ou abaixo do seu valor de mercado.

1.2 A solução

Com a evolução das técnicas de machine learning e da capacidade de processamento dos computadores é possível utilizar diversos métodos científicos para gerar uma ferramenta que precifica com exatidão qualquer imóvel em uma região com um número suficiente de amostras disponíveis para o modelo. Tal produto seria capaz de determinar em segundos qual o preço adequado para um imóvel específico e com baixas margens de erro, ou seja, mais rápido e mais preciso do que os meios convencionais de pesquisa de mercado.

Como base para a ferramenta de precificação de imóveis serão testados os seguintes métodos preditivos: *K Nearest Neighbor* (*KNN*) (FIX, 1989), *Regression* (*linear and logistic*) e *Random Forest* (HO, 1995). Antes da aplicação desses modelos será utilizado o método *Univariate Selection* (BLUM, 1997) para determinar quais são os atributos mais importantes para a precificação de um imóvel, sejam elas: área, localização, quantidade de quartos, tipo de imóvel, facilidades do condomínio, entre outros. Em seguida, após a aplicação dos modelos de machine learning, será escolhido aquele que tiver a melhor generalização do grupo de treinamento para o grupo de teste.

2 REVISÃO BIBLIOGRÁFICA

A fim de facilitar o entendimento do presente trabalho foi realizada a revisão bibliográfica dos seguintes pontos: machine learning, seleção de atributos, aprendizado supervisionado voltado para regressão, investimentos e diferenciais competitivos, panorama de mercado e resultados esperados. Os tópicos são desencadeados de modo a possibilitar a visualização do que se espera obter a partir da aplicação de conceitos teóricos com elementos práticos, materializado na exploração de modelos já realizados e disponíveis no mercado.

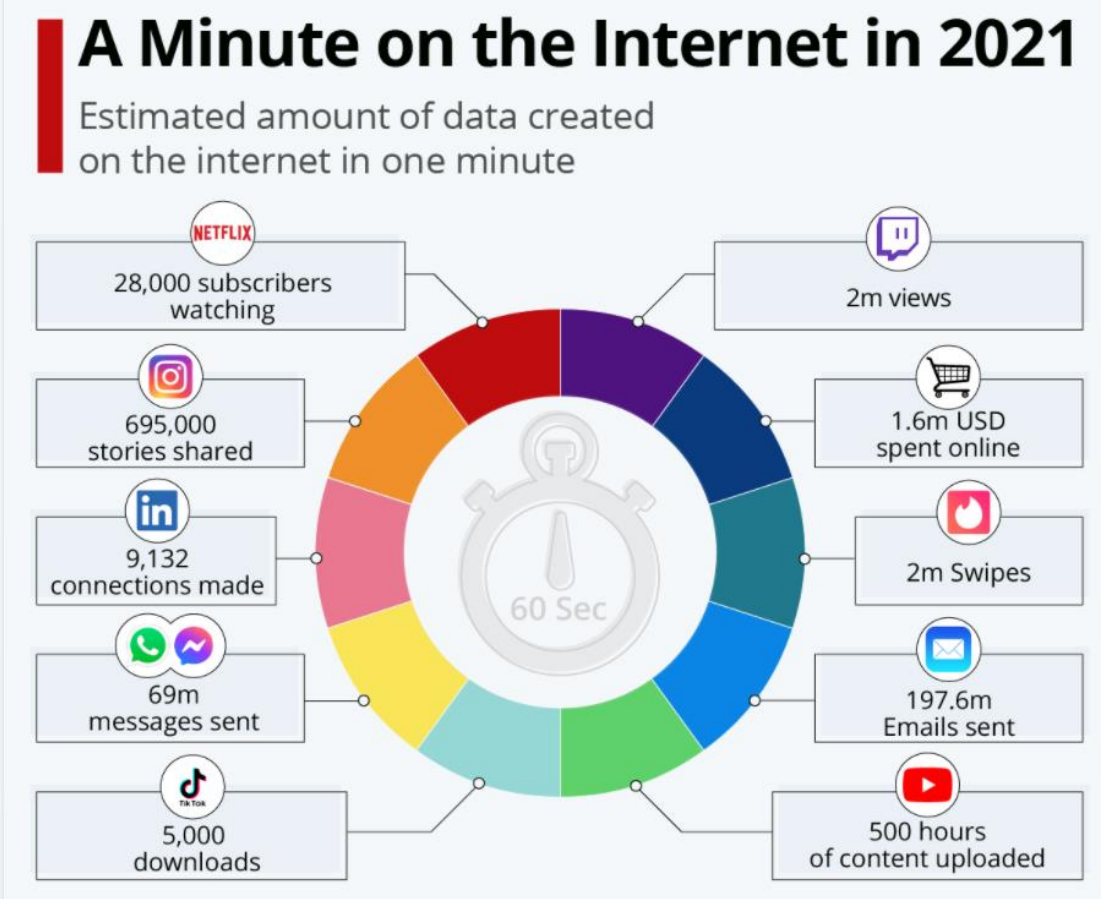
2.1 O conceito de *machine learning*

O *machine learning*, é uma vertente da inteligência artificial a qual utiliza algoritmos computacionais que podem ser aprimorados automaticamente por meio da experiência e do uso de dados (MITCHELL, 1997). Segundo Koza (1996), o *machine learning* utiliza algoritmos para construir um modelo baseado em dados de amostra, conhecidos como "dados de treinamento", a fim de fazer previsões ou decisões sem serem explicitamente programados para isso.

Atualmente, os algoritmos de *machine learning* são usados em uma ampla variedade de aplicações, como na medicina, filtragem de e-mail, reconhecimento de voz e visão computacional, onde é difícil ou inviável desenvolver algoritmos convencionais para realizar as tarefas necessárias (HU, 2020)

Witten, Frank e Hall (2011), em seu livro *Data Mining* afirma: “O que é novo é o aumento desconcertante da possibilidade de encontrar padrões nas informações”. Entretanto, para lidar com tantas possibilidades é necessário a ajuda de máquinas cada vez mais robustas, afinal a cada dia que passa aumenta a quantidade e a complexidade dos dados — estima-se que a quantidade de informações armazenadas no mundo dobre a cada 20 meses o que se torna evidente pela difundida figura 1 que mostra a quantidade de atividades que ocorrem no mundo digital a cada minuto.

Figura 1 – Quantos dados são gerados por minuto nas principais páginas da internet?



Fonte: Lewis (2022).

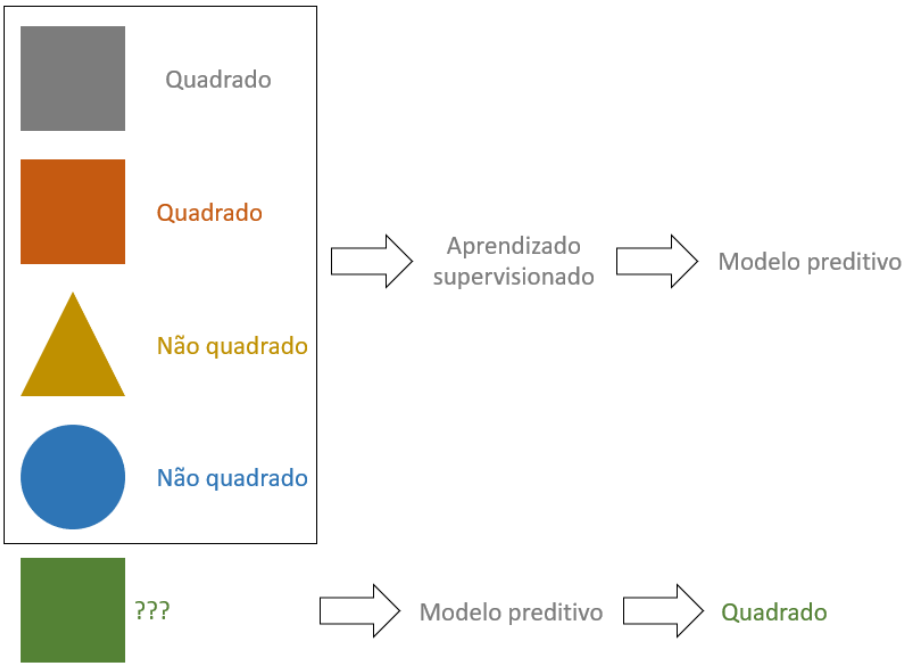
2.2 A seleção de atributos

Já a seleção de atributos é frequentemente usada como um pré-processamento para o aprendizado de máquina. Ela consiste na escolha de um subconjunto de atributos do conjunto original a partir da avaliação de determinados critérios para otimizar o processo de aprendizado. A seleção de características tem sido um campo fértil de pesquisa e desenvolvimento desde 1970 e é comprovadamente eficaz para a remoção de recursos irrelevantes e redundantes, o que: aumenta eficiência nas tarefas de aprendizagem, melhora o desempenho da aprendizagem e aumenta a compreensão dos resultados aprendidos (BLUM, 1997).

2.3 Aprendizado supervisionado voltado para regressão

O aprendizado supervisionado consiste em aprender uma função que mapeia uma entrada para uma saída com base em pares de entrada-saída de exemplo, também chamados de conjunto de exemplos de treinamento. Na aprendizagem supervisionada, cada exemplo é um par que consiste em um objeto de entrada e um valor de saída desejado (MEHRYAR, 2012). No caso da figura 2 as entradas são as imagens de cada animal com seus respectivos rótulos.

Figura 2 – Ilustração do aprendizado supervisionado



Fonte: Elaboração Própria.

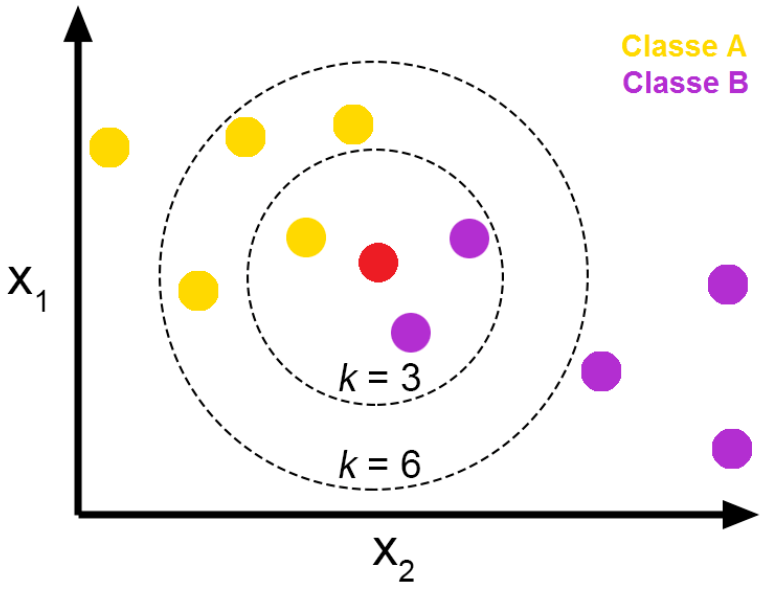
Em um cenário ideal o algoritmo será capaz de rotular corretamente as classes mesmo para instâncias não vistas. Isso requer que o algoritmo de aprendizagem generalize a partir dos dados de treinamento para situações novas de uma forma minimamente assertiva. Essa qualidade estatística de um algoritmo é medida por meio do chamado erro de generalização (CABANNES; BACH; RUDI, 2021)

As técnicas de aprendizado supervisionado que serão utilizadas no projeto serão: *K Nearest Neighbor* (*KNN*), *Linear Regression* e *Random Forest*.

A primeira técnica, *K Nearest Neighbor* (*K-NN*) é um método de classificação não paramétrico desenvolvido pela primeira vez por Evelyn Fix e Joseph Hodges em 1951 (FIX, 1951). Este método é usado para classificação e para regressão. Em ambos os casos, a entrada

consiste nos k exemplos de treinamento mais próximos no conjunto de dados, como mostrado na figura 3, mas a saída depende se *k-NN* é usado para classificação ou regressão:

Figura 3 – Ilustração do método *K Nearest Neightbor* (K-NN)



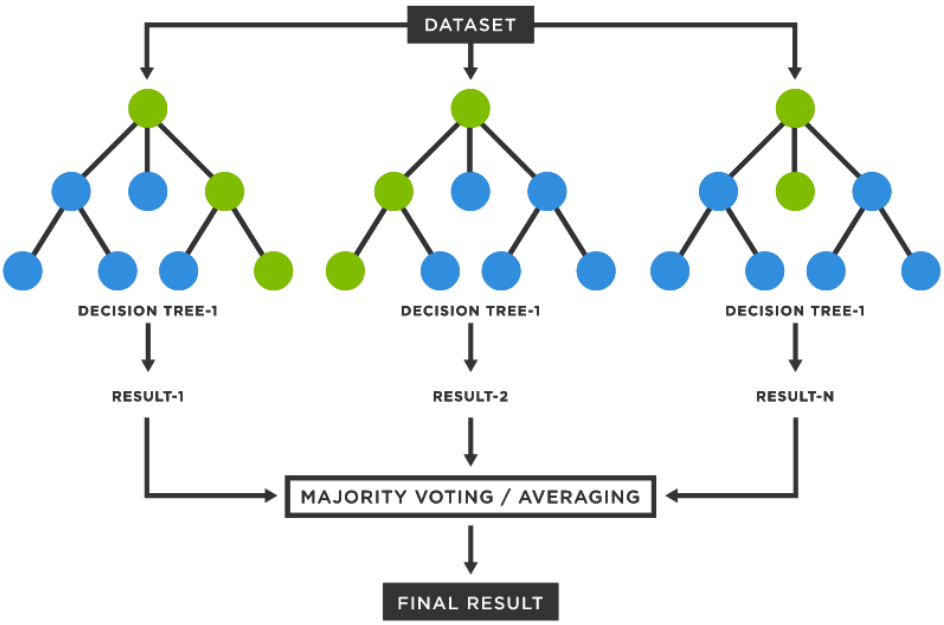
Fonte: Oliveira (2022).

Na classificação *k-NN*, a saída é uma associação de classe. Um objeto é classificado pela ponderação de seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos, sendo k um número inteiro positivo. Se k = 1, então o objeto é simplesmente atribuído à classe daquele único vizinho mais próximo.

Já na regressão *k-NN*, a saída é o valor da propriedade do objeto o qual consiste na média dos valores dos k vizinhos mais próximos.

A segunda técnica utilizada será o *random forests*, que é um método de aprendizagem de conjunto para classificação, regressão e outras tarefas que operam através da construção de uma infinidade de árvores de decisão no momento do treinamento. Para tarefas de classificação, a saída da *random forest* é a classe selecionada pela maioria das árvores, como ilustrado na figura 4. Para tarefas de regressão, a média ou previsão média das árvores individuais é retornada. A grande vantagem das *random forests* frente a outros modelos é que elas corrigem o hábito de ajustar-se em demasia ao seu conjunto de treinamento. O primeiro algoritmo para florestas de decisão aleatória foi criado em 1995 por Tin Kam Ho usando o método do subespaço aleatório, que, na formulação de Ho, é uma forma de implementar a abordagem de "discriminação estocástica" para classificação proposta por Eugene Kleinberg (HO, 1995).

Figura 4 – Ilustração do método *random decision forests*



Fonte: Tibco (2022).

2.4 Panorama de mercado

Em relação a ferramentas similares no mercado, tanto a empresa QuintoAndar quanto a Loft, startups do setor imobiliário que atuam em cidades específicas pelo país, possuem ferramentas online que fazem a avaliação dos imóveis dos proprietários. Porém, nenhuma das duas plataformas permite visões comparativas entre imóveis, o que permitiria a comparação entre custo e benefício para os compradores, principalmente aqueles com perfil de investidor. Da mesma forma, não há hoje disponível em nenhuma plataforma o preço do metro quadrado, valor básico para permitir a comparação entre imóveis em regiões similares. Por exemplo, caso um investidor quisesse adquirir um imóvel em Pinheiros, São Paulo, hoje seria necessária uma comparação árdua entre diversos imóveis disponíveis em diversas plataformas e isto provavelmente teria que ser feito de forma manual, o que levaria bastante tempo e aumenta a suscetibilidade ao erro.

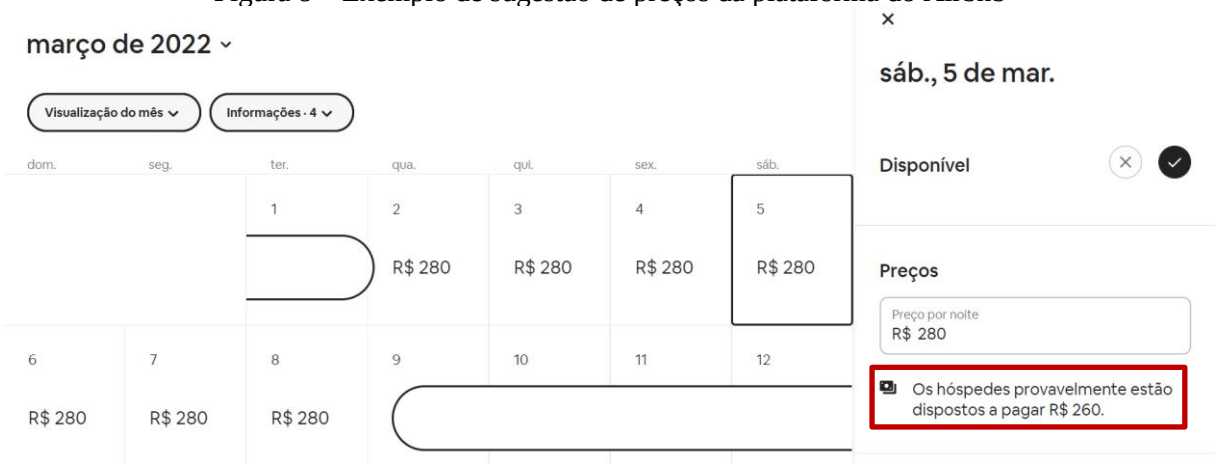
2.4.1 A solução do Airbnb

O Airbnb é uma plataforma digital que conta com milhões de anfitriões e de viajantes que anunciam o seu espaço e reservam alojamentos em qualquer lugar do mundo. Esta plataforma trabalha com aluguéis de curta duração e conta com uma quantidade de informações

disponível que a permite gerar dicas de preços aos novos anfitriões baseadas nas informações de outros imóveis já listados no site assim como baseadas na procura de cada região (AIRBNB, 2021).

Atualmente, a plataforma possui uma ferramenta de dicas de preços e de descontos, o *Smart Pricing*, como é chamado pela plataforma. Esta ferramenta, como ilustrado na figura 8, ajusta automaticamente os preços das diárias com base em mais de 70 fatores que influenciam o preço como: o tipo e a localização do imóvel, temporada, demanda, listagens próximas que foram reservadas, número de avaliações positivas do seu anúncio e outros fatores não revelados pela plataforma (AIRBNB, 2021).

Figura 8 – Exemplo de sugestão de preços da plataforma do Airbnb



Fonte: Airbnb (2021).

Além de poder deixar o controle do preço do aluguel dos imóveis a cargo da plataforma o cliente também sempre no controle do preço e pode substituir essas sugestões a qualquer momento.

Um dos pontos negativos da ferramenta é que ela não levará em consideração tudo aquilo que pode impactar o que os hóspedes estão dispostos a pagar, como uma vista deslumbrante ou o nível de hospitalidade com o qual contribui para a experiência do hóspede. Ou seja, há fatores subjetivos que dependem de uma avaliação humana e que podem impactar os preços de uma estadia, mas que atualmente não são levados em consideração pela ferramenta.

Portanto, para evitar experiências ruins a ferramenta pede que o cliente defina um preço mínimo, para que o preço nunca seja inferior a um valor com o qual o cliente se sente confortável.

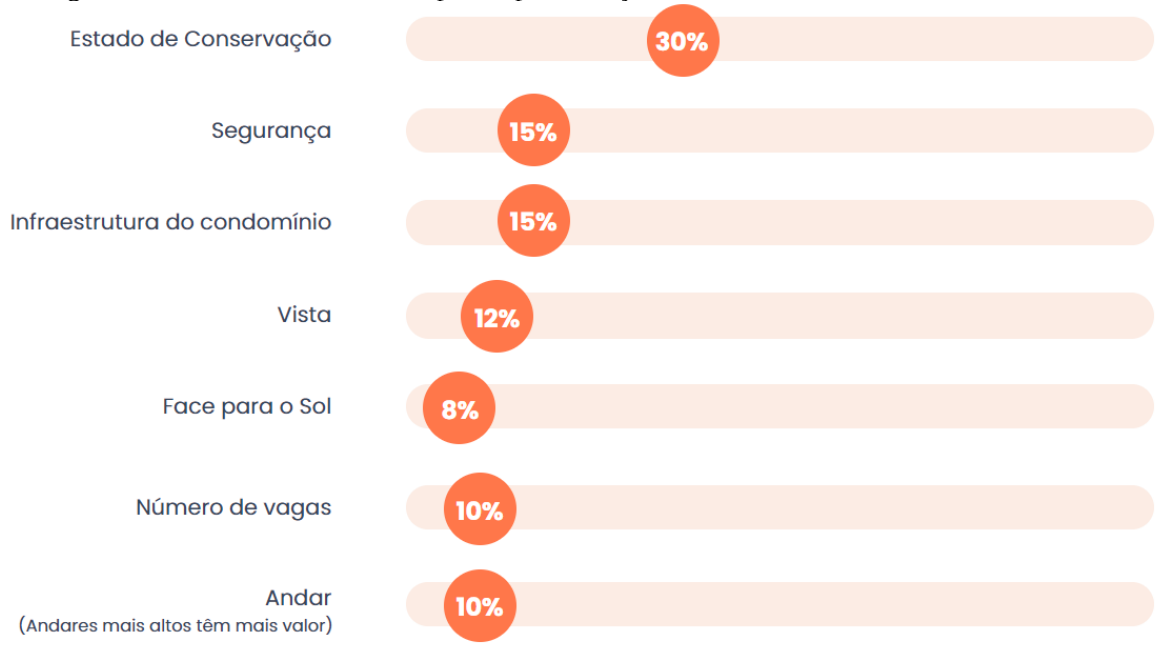
2.4.2 A solução da Loft

De acordo com o site da companhia, a Loft é uma plataforma digital que atua na compra, na venda, na troca e na reforma de apartamentos residenciais. A empresa foi fundada em 2018 e atua em 2021 nas cidades de São Paulo e do Rio de Janeiro (LOFT, 2021).

Diferentemente do Airbnb, a ferramenta de precificação da Loft é voltada para o preço de venda do imóvel e utiliza preços de venda recentes de apartamentos na mesma região e com características similares para gerar um preço sugerido.

Ainda de acordo com a plataforma, os atributos considerados na sugestão de preço são: estado de conservação, segurança, infraestrutura do condomínio, vista, face para o sol, número de vagas e andar, todos ponderados de acordo com a figura 9.

Figura 9 – Atributos e relevância para a precificação do valor de venda de acordo com a Loft



Fonte: Loft (2021).

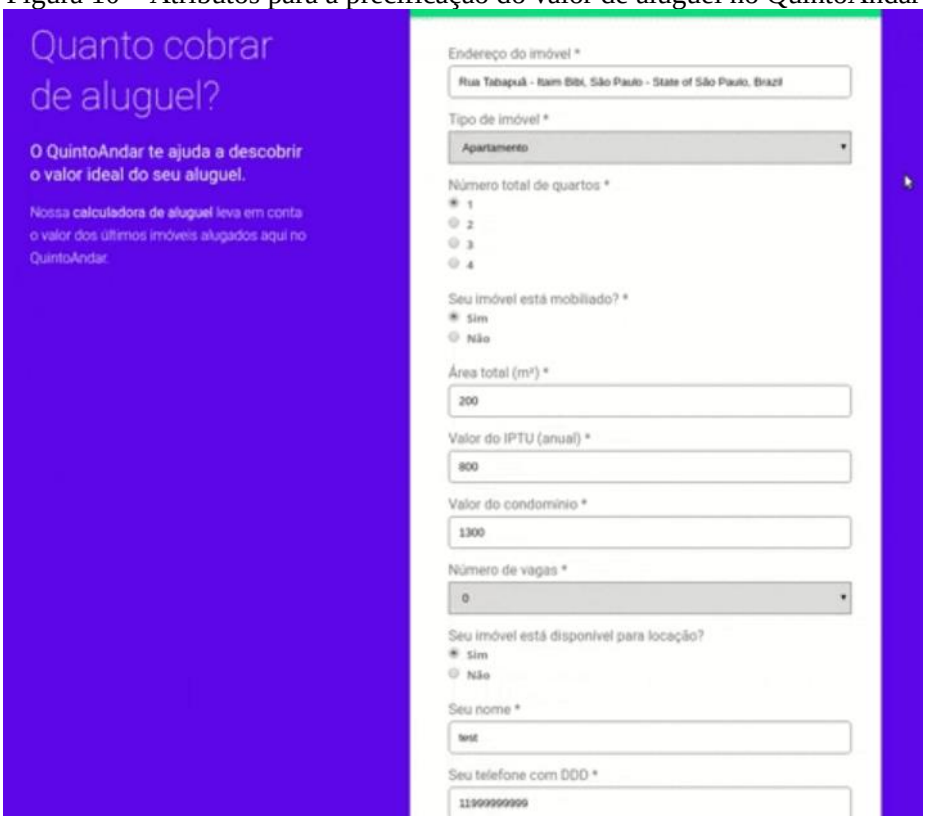
2.4.3 A solução do QuintoAndar

O QuintoAndar é uma imobiliária digital brasileira que atua no segmento de: compra, venda e aluguel de imóveis. A plataforma se diferencia das imobiliárias tradicionais reduzindo a burocracia ao não exigir fiador ou cheque caução ao mesmo tempo que mantem uma garantia de pagamento ao proprietário das mensalidades e de possíveis danos ao imóvel (QUINTOANDAR, 2021).

Assim como o Airbnb e a Loft, o QuintoAndar possui uma grande quantidade de informações disponíveis sobre milhões de imóveis no Brasil o que permite com que a plataforma ofereça a Calculadora QuintoAndar, ferramenta da imobiliária digital que ajuda os proprietários a definir os valores de aluguel.

De acordo com Quintoandar (2021), “A ferramenta faz um estudo, também por meio de inteligência artificial, em uma base de dados com 30 especificidades dos imóveis já anunciados e indica ao proprietário o melhor valor para anunciar o seu bem. Levando em consideração a cifra média de imóveis semelhantes.”.

Figura 10 – Atributos para a precificação do valor de aluguel no QuintoAndar



Fonte: Barbosa (2019).

Como observado na figura 10, que mostra o formulário necessário para calcular o preço de aluguel do imóvel algumas especificidades que são levadas em conta são: endereço, tipo de imóvel (apartamento, kitnet, casa ou casa em condomínio), quantidade de quartos, ar condicionado, área (m²), valor do condomínio e quantidade de banheiros.

Como resultado a plataforma disponibiliza um intervalo de preços sugeridos como mostrado na figura 11. Em contrapartida, a plataforma consegue gerar um volume de pessoas possivelmente interessadas em utilizar seus serviços gerando mais tráfego para o seu website e consequentemente aumentando seu volume de clientes.

Figura 11 – Resultado da ferramenta disponibilizada pelo QuintoAndar

Fonte: Barbosa (2019).

2.5 Investimentos e diferenciais competitivos

Para a construção do produto mínimo viável, ou MVP, será necessária uma base de dados representativa de uma cidade em qualquer lugar do planeta contendo as informações que forem identificadas como fundamentais para a atribuição do valor de um imóvel, sejam elas: bairro, área, quantidade de quartos, tipo, facilidades, entre outros. Para conseguir esta base de dados serão utilizadas inicialmente ferramentas de busca, mas é possível que seja necessária a utilização de ferramentas de *web scraping*.

Após a aquisição das bases de dados será necessário investir tempo na construção dos algoritmos para processá-la e gerar os resultados nos grupos de teste. Em seguida, será possível analisar o resultado e a capacidade de generalização do modelo permitindo a escolha do algoritmo com melhor capacidade de generalização.

Análise SWOT

- Forças: inovação para os investidores, produto único no mercado (comparador de imóveis), escalabilidade, praticidade, acessibilidade, replicabilidade para outras localidades;
- Fraquezas: facilmente replicável, depende de dados de terceiros;
- Ameaças: concorrência (outras plataformas já disponibilizam calculadora para precificar imóveis e podem também disponibilizar comparadores de imóveis);
- Oportunidades: monetizar a venda de leads de imóveis, monetizar via anúncios, monetizar via venda de plataforma profissional de comparação de imóveis;

2.6 Resultados esperados

Dentre os resultados esperados, destacam-se:

- Precificação de imóveis em qualquer cidade do país baseada em dados dinâmicos de diversas plataformas distintas de anúncio de imóveis. A inovação viria da capacidade dinâmica de replicar o modelo também para cidades de médio e de pequeno porte.

Dentre os produtos derivados deste projeto, destacam-se:

- Monetizar os leads adquiridos por meio desta ferramenta para outras plataformas especializadas em compra e venda de imóveis, propósito para o qual ainda não foram encontrados concorrentes.
- Criação de site capaz de acessar múltiplas plataformas e com capacidade de capturar imóveis com o valor do metro quadrado abaixo do preço médio do mercado, sinalizando uma possível oportunidade de compra. Este projeto seria completamente inovador, já que no mercado não há ferramentas que comparem rapidamente imóveis e muito menos que sinalizem oportunidades baseada no preço do metro quadrado inferior ao preço de mercado.

3 METODOLOGIA E DESENVOLVIMENTO DE MODELOS DE PREDIÇÃO DE PREÇOS DE VENDA DE IMÓVEIS

Este capítulo tem por objetivo apresentar as etapas do desenvolvimento deste trabalho. A seção 3.2 irá introduzir o processo de obtenção dos dados. A seção 3.3 tratará do pré-processamento de dados, assim como seus objetivos. Já a última seção deste capítulo, 3.4, apresentará os resultados da predição dos preços dos imóveis com base nos modelos de regressão *random forest*, *KNN* e linear multivariada.

3.1 Obtenção dos Dados

Os dados necessários para a criação de um modelo de predição de preços de imóveis requer uma quantidade de imóveis representativa dentro de determinada cidade. Porém, também é necessário que esta base contenha o máximo de atributos possíveis destes imóveis, como: quantidade de quartos, quantidade de banheiros, quantidade de metros quadrados, localização, dentre outros.

Para este trabalho foi utilizada a base de imóveis de Melbourne[1] com dados de setembro de 2017 extraídos de informações de imóveis publicamente disponíveis na página domain.com.au. O conjunto de dados inclui as seguintes informações:

- *Suburb*: bairro;
- *Address*: endereço;
- *Rooms*: número de quartos;
- Price: preço em dólares;
- Tipo: br - quarto(s); h - casa, chalé, vila, semi, terraço; u - unidade, duplex; t - casa geminada; dev site - site de desenvolvimento; o res – outros imóveis residenciais;
- *SellerG*: agente imobiliário;
- *Date*: data de venda;
- *Distance*: distância do centro da cidade;
- *Regionname*: Região geral (Oeste, Noroeste, Norte, Nordeste...etc);

[1] Disponível em: < https://www.kaggle.com/anthonypino/melbourne-housing-market com> Acesso em: 01/05/2022

- *Propertycount*: número de propriedades que existem no subúrbio;
- *Bedroom2*: nº de quartos;
- *Bathroom*: número de banheiros;
- *Car*: número de vagas;
- *Landsize*: tamanho do terreno;
- *BuildingArea*: tamanho do terreno;
- *BuildingArea*: tamanho do edifício;
- *CouncilArea*: conselho de administração para a região.
- *Lattitude*: latitude;
- *Longtitute*: longitude;

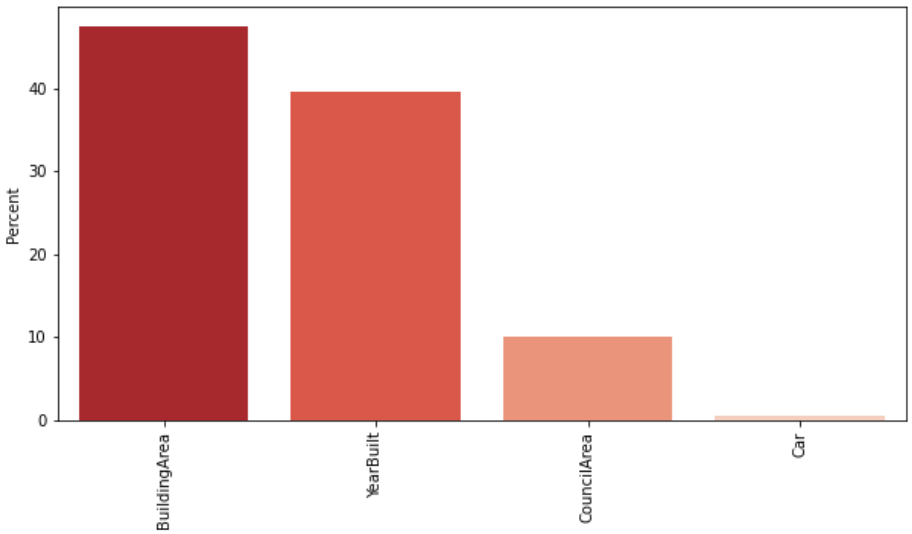
São ao todo 13.580 imóveis com 19 atributos, sendo ao todo 6 atributos de texto, 11 atributos numéricos, 1 atributo de data de publicação e 1 atributo alvo, o preço de venda.

3.2 Pré-processamento de Dados

O primeiro passo para tornar os dados utilizáveis pelos algorítmos de *data science* foi identificar a quantidade de valores de atributos não disponíveis dentre os 13.580 imóveis e eliminá-los. Ao eliminar os imóveis com informações faltantes restaram 6.196 imóveis, ou 45,6%.

Figura 13 – Quantidade de valores indisponíveis por atributo

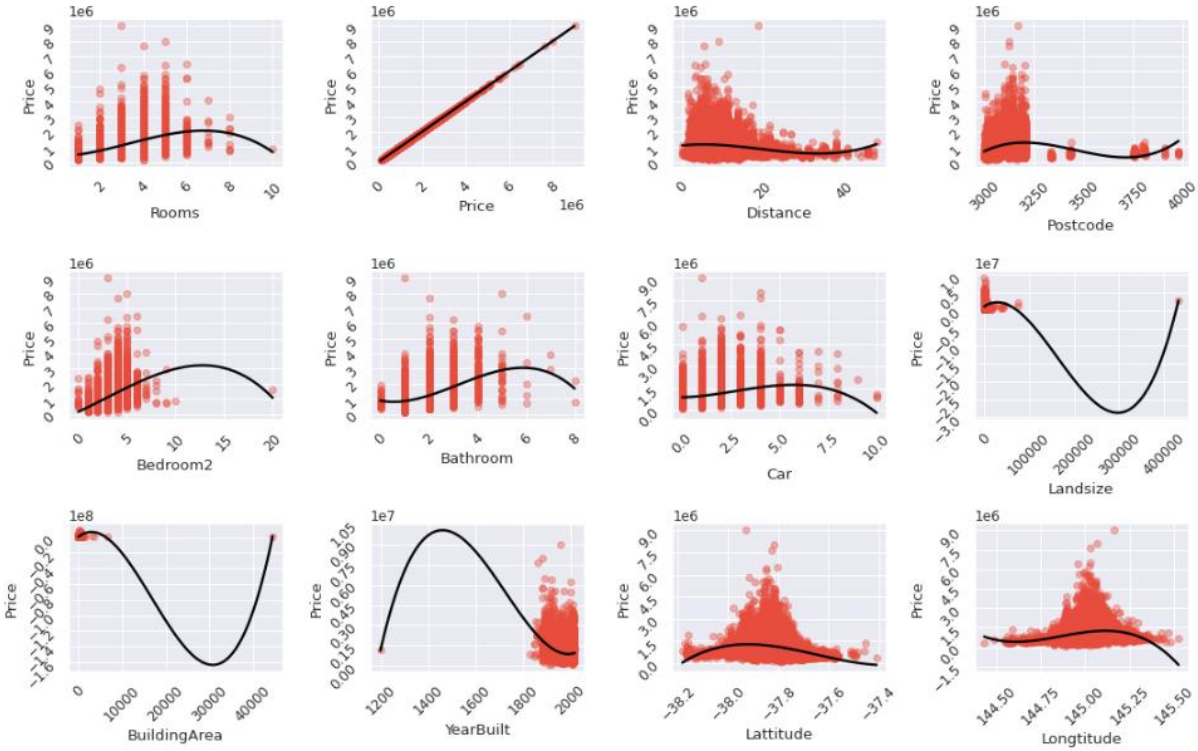
	BuildingArea	YearBuilt	CouncilArea	Car
Total	6450.000000	5375.000000	1369.000000	62.000000
Percent	47.496318	39.580265	10.081001	0.456554



Fonte: Elaboração própria a partir de dados do Kaggle (2022)

O passo subsequente foi identificar os *outliers* e removê-los de forma a otimizar os resultados dos modelos preditivos. De acordo com Rodrigues (2017), por meio da observação gráfica conjugada a comparação de modelos de reta de regressão é possível determinar se é necessária ou não a exclusão dos mesmos. Portanto, foram preparados os gráficos na figura 14 no intuito de definir valores limites superiores e inferiores que mantivessem o maior número possível de amostras, mas que também eliminassem pontos demasiadamente fora da curva de tendência principal.

Figura 14 – Detecção de outliers por meio de observação gráfica



Fonte: Elaboração própria a partir de dados do Kaggle (2022)

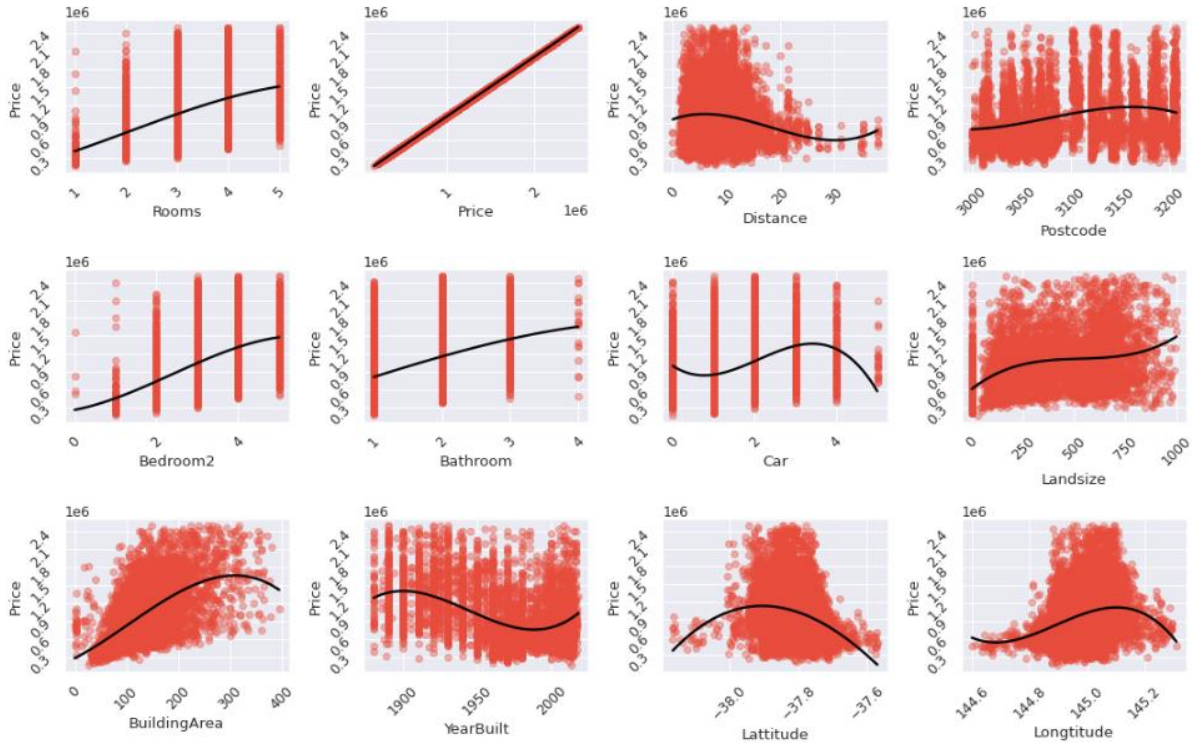
Com base na observação da figura 14, as seguintes remoções foram realizadas e contribuíram para reduzir o erro médio das predições:

- *Rooms*: > 5;
- *Price*: > $2.500.000,00;
- *Regionname*: diferente de *Eastern Victoria*, *Northen Victoria* e *Western Victoria*.
- *Bedroom2*: > 5;
- *Car*: > 5;

- *Bathroom*: > 4;
- *Postcode*: : 3207;
- *BuildingArea*: > 400;
- *YearBuilt*: < 1880;
- Landsize: > 1000;
- *SquaredMetersPrice*: > $20.000,00;

Como resultado restaram 5.510 imóveis, ou 40.6% dos imóveis e o número de *outliers* foi reduzido de forma a permitir melhores ajustes das curvas médias, como visto na figura 15:

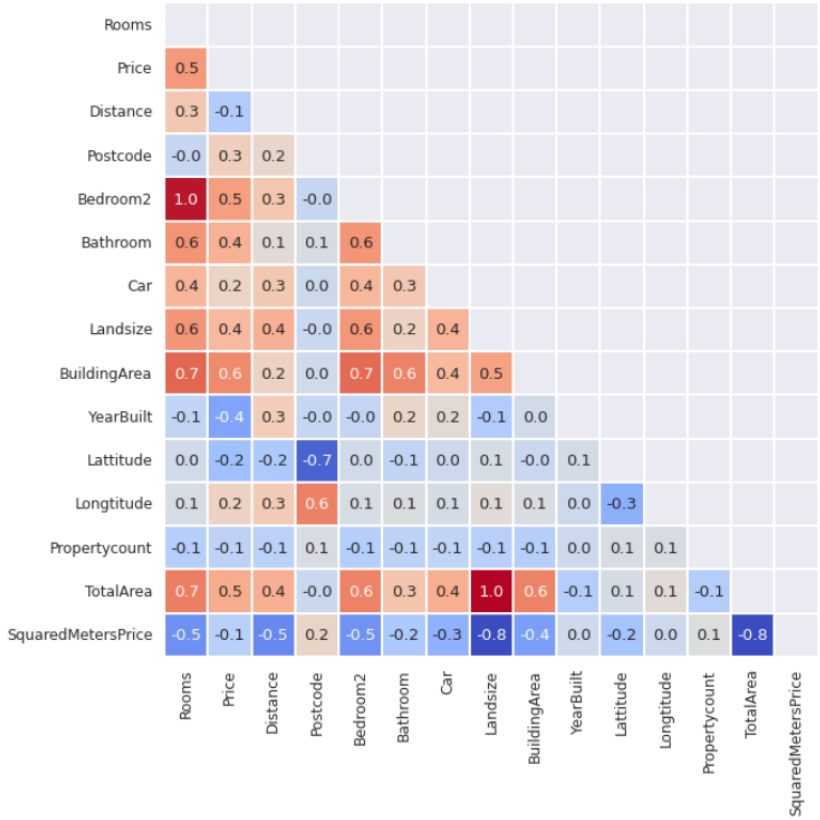
Figura 15 – Resultado pós-elminiação de *outliers*



Fonte: Elaboração própria a partir de dados do Kaggle (2022)

Por último, a correlação entre os atributos numéricos utilizando o método de Pearson (FIGUEIREDO FILHO, 2009) foi levantada mostrando que, de acordo com a figura 16. os cinco atributos que mais tem impacto no preço dos imóveis são: *BuildingArea* +0,6, *'Bedroom2'* +0,5, *'Bathroom'* +0,4, *'Landsize'* +0,4, *'YearBuilt'* -0,4.

Figura 16 – Correlação entre os atributos numéricos



Fonte: Elaboração própria a partir de dados do Kaggle (2022)

3.3 Resultados

3.3.1 Resultados utilizando o modelo preditivo *Random Forest*

Inicialmente foi realizado o teste com os atributos com maior correlação servindo de entrada para o modelo *Random Forest* e o resultado foi:

Rodada com os cinco atributos com maior correlação com o preço do imóvel e com remoção de linhas com valores indisponíveis:

a. Atributos = [*'BuildingArea'*, *'Bedroom2'*, *'Bathroom'*, *'Landsize'*, *'YearBuilt'*]
b. Erro médio absoluto percentual: 24,69%;
c. Erro médio absoluto: $269.547,00;
d. Desvio padrão: $457.123,30;
e. Amostras: 6.196 sendo 4.647 (75%) treino e 1549 (25%) validação.

O resultado foi um erro de 24,69% comparando as predições aos preços publicados na plataforma. Portanto, uma nova rodada de testes foi executada adicionando mais atributos ao modelo preditivo.

Rodada com onze atributos, com remoção de linhas com valores indisponíveis mas sem remoção de outliers:

a. Atributos = [*'Rooms'*, *'Distance'*, *'Postcode'*, *'Bedroom2'*, *'Bathroom'*, *'Car'*, *'Landsize'*, *'BuildingArea'*, *'YearBuilt'*, *'Lattitude'*, *'Longtitude'*]
b. Erro médio absoluto percentual: 15,11%;
c. Erro médio absoluto: $173.454,00;
d. Desvio padrão: $355.019,30;
e. Amostras: 6.196 sendo 4.647 (75%) treino e 1549 (25%) validação.

O resultado teve uma melhora significativa de um erro percentual de 24,69% para 15,11%, ou -9,58 pontos percentuais. Já o erro absoluto teve uma melhora de $269.547,00 para $173.454,00 o que representa um ganho de -35,64% no erro. Por último o desvio padrão do erro de predição também foi reduzido de $457.123,30 para $355.019,30, uma otimização de -22,33%.

Em uma última tentativa de otimizar os resultados foram removidos os valores outliers conforme demonstrado na seção 3.3 sobre o pré-processamento dos dados.

Rodada com onze atributos, com remoção de linhas com valores indisponíveis e com remoção de outliers:

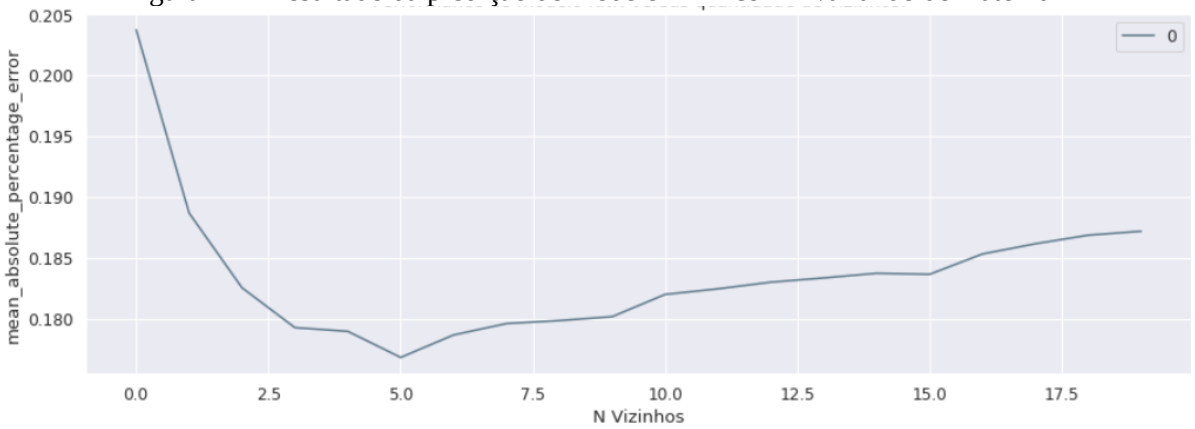
a. Atributos = [*'Rooms'*, *'Distance'*, *'Postcode'*, *'Bedroom2'*, *'Bathroom'*, *'Car'*, *'Landsize'*, *'BuildingArea'*, *'YearBuilt'*, *'Lattitude'*, *'Longtitude'*]
b. Erro médio absoluto percentual: 13.34%;
c. Erro médio absoluto: $134.663,00;
d. Desvio padrão: $190.456,30;
e. Amostras: 5.510 sendo 4.132 (75%) treino e 1.378 (25%) validação

Os novos resultados não são comparáveis de forma justa aos resultados anteriores já que o conjunto de treino e de teste foram alterados com a remoção dos *outliers*. Porém, a intenção de removê-los teve seu objetivo cumprido que era reduzir o erro percentual, o erro absoluto e o desvio padrão. Desta forma, os testes com os demais modelos serão realizados utilizando a base de dados removendo as amostras destoantes.

3.3.2 Resultados utilizando o modelo preditivo *KNN*

Para realizar a predição no modelo *KNN* foram utilizados os dados do último e melhor cenário do modelo *Random Forest*, ou seja, onze atributos com remoção de linhas com valores indisponíveis e com remoção de outliers. Além disso, foi realizado um teste com o número de vizinhos variando de um até vinte para identificar qual seria o ponto ótimo deste modelo.

Figura 17 – Resultado da predição do modelo *KNN* com n variando de 1 até 20



Fonte: Elaboração própria a partir de dados do Kaggle (2022)

Como é possível observar na figura 17, o resultado ideal ocorreu com N = 5 em rodada com onze atributos, com remoção de linhas com valores indisponíveis mas sem remoção de outliers:

a. Atributos = ['Rooms', 'Distance', 'Postcode', 'Bedroom2', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'YearBuilt', 'Lattitude', 'Longtitude']
b. Erro médio absoluto percentual: 17,68%;
c. Erro médio absoluto: $173.116,00;
d. Desvio padrão: $240.138,80;
e. Amostras: 5.510 sendo 4.132 (75%) treino e 1.378 (25%) validação

O resultado utilizando o modelo *KNN* com n igual a 5 atingiu um erro percentual de 17,68% ou 4,34 pontos percentuais a mais que os 13,34% do modelo *Random Forest*. O erro absoluto foi de $173.116,00 versus $134.663,00 o que representa uma piora de +28,55% no erro. Por último, o desvio padrão do erro de predição também foi maior atingindo $240.138,80 versus $190.456,30, uma piora de +26,09%.

3.3.3 Resultados utilizando o modelo preditivo Regressão Linear Multivariada

Para realizar a predição no modelo Regressão Linear Multivariada foram utilizados os dados do último e melhor cenário do modelo *Random Forest*, ou seja, onze atributos com remoção de linhas com valores indisponíveis e com remoção de outliers.

a. Atributos = ['Rooms', 'Distance', 'Postcode', 'Bedroom2', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'YearBuilt', 'Lattitude', 'Longtitude']
b. Erro médio absoluto percentual: 21,37%;
c. Erro médio absoluto: $193.697,00;
d. Desvio padrão: $252.759,70;

e. Amostras: 5.510 sendo 4.132 (75%) treino e 1.378 (25%) validação

O resultado utilizando o modelo Regressão Linear atingiu um erro percentual de 21,37% ou 8,03 pontos percentuais a mais que os 13,34% do modelo *Random Forest*. O erro absoluto foi de $193.697,00 versus $134.663,00 o que representa uma piora de +43,84% no erro. Por último, o desvio padrão do erro de predição também foi maior atingindo $252.759,70 versus $190.456,30, uma piora de +32,71%.

4 CONCLUSÃO

Este trabalho apresentou um projeto de negócio envolvendo o uso de modelos de *machine learning* para calcular o valor de venda de imóveis.

Foram utilizados os métodos *Random Forest*, *KNN* e regressão linear multivariada. Os resultados foram: para o modelo *Random Forest* com remoção de outliers um erro percentual de 13,34%, um erro absoluto de $134.663,00 e um desvio padrão do erro de $190.456,30; para o modelo *KNN* com remoção de outliers um erro percentual de 17,68% , um erro absoluto de $173.116,00 e um desvio padrão do erro de $240.138,80; para o modelo Regressão Linear Multivariada com remoção de outliers um erro percentual de 21,37%, um erro absoluto de $193.697,00 e um desvio padrão do erro de $252.759,70.

Portanto, o modelo de melhor desempenho que seria escolhido para um mínimo produto viável deste projeto seria o modelo *Random Forest* com remoção de outliers e erro percentual de 13.34% na predição.

Em trabalhos futuros, considerar:

- O erro entre o valor predito e o valor da plataforma pode ser causado pela própria má precificação dos proprietários;
- A inclusão de mais amenidades que agregam valor, como: vista privilegiada, móveis planejados, piscina, painéis solares, dentre outros aspectos que valorizam imóveis;
- A inclusão de variáveis de localização como: segurança, proximidade de estações de metrô e de shoppings, dentre outros.
- A necessidade de atualizações períodicas para considerar aspectos como inflação e como valorização regional.
- A criação da clusterização dos imóveis de cada região em: acima do preço e abaixo do preço de mercado utilizando o produto desenvolvido associado a modelos de classificação.

REFERÊNCIAS BIBLIOGRÁFICAS

AGÊNCIA CBIC (2021). **Brasileiros querem investir mais em imóveis.** Portal da Agência CBIC [online]. Disponível em: <https://cbic.org.br/brasileiros-querem-investir-mais-em-imoveis/>. Acesso em: 26 set. 2021.

AIRBNB (2021). **O que é o Airbnb e como ele funciona?** Portal Airbnb [online]. Disponível em: <https://www.airbnb.com.br/help/article/2503/o-que-%C3%A9-o-airbnb-e-como-ele-funciona>. Acesso em: 25 dez. 2021.

BARBOSA, A. (2019). **Machine Learning Pipeline at QuintoAndar.** Portal QuintoAndar Tech Blog [online]. Disponível em: <https://medium.com/quintoandar-tech-blog/machine-learning-pipeline-at-quintoandar-e2f24136006b>. Acesso em: 08 jan. 2022.

BLUM, A. L. (1997). **Selection of Relevant Features and Examples in Machine Learning**. Palo Alto (EUA): Pat Langley, 1997.

CABANNES, V.; BACH, F.; RUDI, A. (2021). **Fast Rates for Structured Prediction**. Boulder (EUA): Colt, 2021.

FIGUEIREDO FILHO, D. B. (2009). Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Política Hoje**, v. 18, n. 1, Recife (PE), p. 115-146.

FIX, E.; HODGES, J. L. (1981). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. **International Statistical Review,** vol. 57, n. 03, 1989, p. 238-247.

HO, T. K. (1997). Random Decision Forests. *In*: **Proceedings on the Third International Conference on Document Analysis and Recognition, vol. 1, Montreal (Canadá), 1997,** p. 278-282.

HU, J.; NIU, H.; CARRASCO, J.; LENNOX, B.; ARVIN, F. (2020). Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. **IEEE Transactions on Vehicular Technology,** vol. 69, n. 12, 2020, p. 14413-14423.

KOZA, J. R.; BENNETT III, F. H.; ANDRE, D.; KEANE, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. *In:* GERO, J. S.; SUDWEEKS, F. (org). **Artificial Intelligence in Design '96**. Nova Iorque (EUA): Springer, 1996. p. 151-170.

LOFT (2021). **Descubra o valor do seu apê.** Portal LOT [online]. Disponível em: <https://mkt.loft.com.br/calculadora-valor-apartamento>. Acesso em: 25 dez. 2021.

MACQUEEN, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *In*: **Fifth Berkeley Symposium - MacQueen on mathematical statistics and probability**, 1967, Berkeley (EUA), p. 281-297. Disponível em: <https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf>. Acesso em: 29 maio 2022.

MITCHELL, T. (1997). **Machine Learning**. Nova Iorque (EUA): McGraw Hill, 1997.

OLIVEIRA, I. J. G. **KNN (K-Nearest Neighbors) #1**. Portal AiBrasil [online]. Disponível em: <https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e>. Acesso em: 08 jan. 2022.

QUINTOANDAR (2021). **Saiba como a tecnologia pode te ajudar a investir em imóveis**. Portal QuintoAndar [online]. Disponível em: <https://conteudos.quintoandar.com.br/investimento-em-imoveis-como-a-tecnologia-pode-te-ajudar/>. Acesso em: 25 dez. 2021.

RODRIGUES, E.C. (2017). **Modelos de Regressão Múltipla.** Portal UFOP [online]. Disponível em: < http://professor.ufop.br/sites/default/files/ericarodrigues/files/regressaolinearmultipla_parte8.pdf>. Acesso em: 05 sep. 2021.

TIBCO (2022). **What is a Random Forest?** Portal TIBCO [online]. Disponível em: <https://www.tibco.com/reference-center/what-is-a-random-forest>. Acesso em: 08 jan. 2022.

WITTEN, I. H.; FRANK, E.; HALL, M. A. (2011). **Data Mining**: **practical machine learning tools and techniques**. 3. ed. Burlington (EUA): Morgan Kaufmann, 2011.