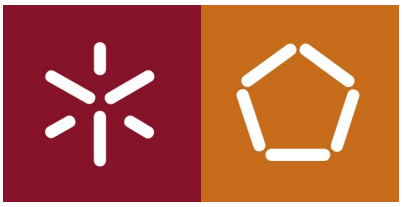


Universidade do Minho
mestrado integrado em engenharia informática

Laboratórios de Engenharia Informática

(2º Semestre / 4º Ano)

ScrapMyProp

Diogo Braga (a82547)
Ricardo Caçador (a81064)
Rui Ribeiro (a80207)

Conteúdo

1 Introdução

O **ScrapMyProp** (Scrap My Property) é um projeto que surge da necessidade de auxiliar qualquer cidadão em compras imobiliárias, de forma a que qualquer investimento efetuado seja o máximo rentável possível.

O mercado de valores imobiliário possui subidas e descidas consoante as qualificações das diferentes zonas distribuídas pelo país, tal como é conhecido. Normalmente, os cidadãos, em alguma fase da sua vida, têm interesse em comprar uma residência e, para tal, utilizam redes imobiliárias para satisfazer as suas necessidades. Para além da intenção de qualquer cidadão querer pagar o valor real do imóvel e assim diminuir a margem de lucro da empresa imobiliária, um cidadão preocupa-se, de igual forma, com a variação de preço que a sua residência vai sofrer após efetuada a compra e, nesse sentido surge esta aplicação, que no nosso caso terá em conta o espaço geográfico do munícipio de Braga.

Utilizando o ScrapMyProp, o processo tem menos probabilidade de ser influenciado para benefício de terceiros. Inicialmente, são identificados os terrenos e os imóveis presentes na Internet no espaço geográfico no qual surge o interesse. De seguida, através de programas de *web-crawling* e *web-scraping*, são explorados casos com características semelhantes ao imóvel que se procura. Por fim, são realizadas técnicas de *machine learning* para classificação e previsão do preço do imóvel. Desta forma é possível calcular o menor valor possível dum imóvel e prever as tendências de preço que este terá.

Quanto à estrutura do relatório, no segundo capítulo é apresentada a motivação para este projeto enquanto no terceiro é explorada a fundamentação teórica. No quarto capítulo é realizado o estado de arte, sendo que, no quinto, é apresentado um levantamento de requisitos do sistema. No sexto capítulo é apresentado o primeiro dos objetivos do projeto, isto é, a recolha de dados e informação através de macanismos de *webscraping*. No sétimo capítulo é apresentado o pré-processamento e análise dos dados recolhidos. Após análise dos dados, são aplicadas técnicas e algoritmos de regressão no capítulo oito. No capítulo nove é realizada a análise da regressão efetuada, dando por concluída o segundo dos objetivos do projeto. No décimo capítulo são aplicadas técnicas de previsão do *target*. Aliado a isso, no capítulo onze é apresentada a análise preditiva, dando por concluído o terceiro dos três objetivos do projeto. No último capítulo é apresentada a conclusão do projeto.

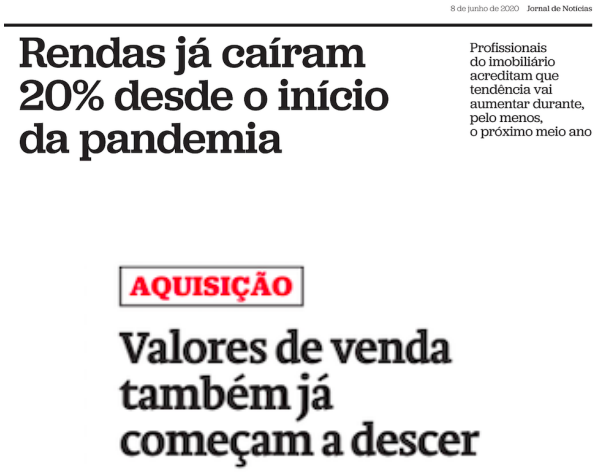
2 Motivação

O tema deste projeto refere-se a algo cuja necessidade existirá, quase sempre, na vida de um cidadão. Seja para realizar um investimento ou apenas para adquirir a futura casa, a consulta do mercado imobiliário é um assunto delicado e que deve ser feito utilizando o máximo de recursos possíveis que auxiliem o cliente na escolha do melhor imóvel pelo melhor preço.

Tendo em conta a altura relativa à data do projeto em que uma pandemia influenciou toda a vida de uma população torna-se ainda mais motivador tentar e aproximar a previsão das subidas e descidas dos valores dos imóveis. Esta pandemia que se iniciou em Portugal, pelos dados oficiais, no dia 2 de Março de 2020 causava, segundo o *Jornal de Notícias* uma quebra de 20%, em média. nas rendas das casas de todo o país e que a tendência seria o aumento desta quebra.

Estranhamente, apesar dos especialistas terem indicado o aumento da tendência de quebra para o próximo meio ano, segundo o *O Minho*, o preço das casas do concelho de Braga aumentou 3,1% desde o início da pandemia até à data da notícia.

Por esta necessidade de avaliar imóveis e pela existência de fatores externos raros como é o caso de uma pandemia, este projeto pretende tornar essa avaliação mais fácil e fornecer uma alternativa à avaliação de um preço de um imóvel que seja corrente e útil para os próximos tempos, funcionando também como um concorrente para o único mecanismo existente para este efeito, o do *Ministério das Finanças* que tem em conta outros fatores que não os das imobiliárias.


8 de junho de 2020 Jornal de Notícias

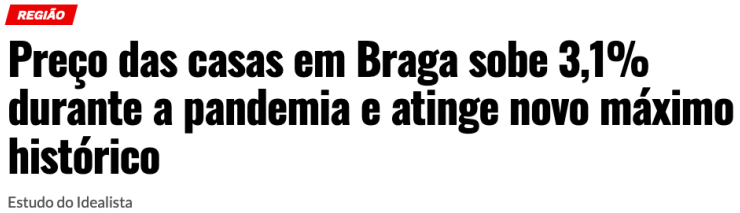
Rendas já caíram 20% desde o início da pandemia

Profissionais do imobiliário acreditam que tendência vai aumentar durante, pelo menos, o próximo meio ano

AQUISIÇÃO

Valores de venda também já começam a descer

Figura 1: Notícias do *Jornal de Notícias* no dia 8 de Junho de 2020


REGIÃO

Preço das casas em Braga sobe 3,1% durante a pandemia e atinge novo máximo histórico

Estudo do Idealista

Figura 2: Notícia do site *O Minho* no dia 1 de Julho de 2020

3 Fundamentação teórica

Para o software ser capaz de auxiliar o utilizador em compras imobiliárias, este tem que ser capaz de avaliar o imóvel em causa. Neste sentido, é natural que a **avaliação imobiliária** seja uma das áreas de estudo abordadas nesta secção.

3.1 Avaliação imobiliária

O setor imobiliário, que engloba uma grande variedade de transações de edifícios e terrenos, distingue-se de outros setores de investimento por razões como: [2]

- Mobilização de fundos avultados;
- Tendência para a valorização da propriedade com o tempo, acima da inflação;
- Propriedade objeto de interferência por medidas políticas e administrativas, que, associadas ao longo prazo de investimentos, tende a aumentar a incerteza e o risco.

Em parte devido a estas menções, o setor imobiliário é uma área de investimento muito apetecível pois é possível, tanto desde uma simples compra e venda até à aquisição de bens para acrescentar mais-valias, rentabilizar um investimento. A grande ocorrência de investimentos conduziu à necessidade de se estudarem formas de identificar o valor dos bens transacionados para lidar com a infinidade de situações que podem acontecer.

3.1.1 Objetivos da avaliação

No contexto imobiliário, o conceito de avaliação pode ser definido como:

Um processo que implica a inspeção de um avaliador a um imóvel, com o propósito de estimar o valor financeiro do mesmo. A avaliação de um imóvel é um processo habitualmente requerido sempre que está em causa a transação de um apartamento, moradia, loja ou terreno, para efeitos de venda, arrendamento ou no âmbito de um crédito à habitação. [3]

O processo de avaliação não é separado dos intervenientes no mercado, o que leva a que sejam eles a fixar o preço dos bens imobiliários. Assim, a avaliação do valor do investimento imobiliário não é apenas baseada no mercado, mas também nas decisões dos investidores individuais, que refletem as estimativas subjetivas de fatores relevantes.

Um avaliador pode ser solicitado a dar a sua estimativa de valor sobre propriedades de tipos diferentes e para finalidades distintas. Dada a variedade de situações, as técnicas utilizadas em certos casos podem não ser eficazes sempre. Por consequência, foram desenvolvidas várias abordagens distintas, que originaram vários métodos de avaliação.

3.1.2 Critérios de valorização

Qualquer avaliação deve ser realizada com prudência e veracidade, visto que podem existir terceiros a tentar inflacionar os valores envolvidos. Desta forma, existem alguns princípios que devem ser seguidos na valorização dos imóveis:

- **Princípio do maior e melhor uso:** Quando se avalia um imóvel, nem sempre este está afeto relativamente ao uso que maximiza o seu valor. Desta forma, o valor de um imóvel, apesar de poder albergar diferentes usos, costuma ser o que resulta do seu uso mais provável, mas tentando também considerar qual o maior retorno possível.

- **Princípio de substituição:** O valor de um imóvel é sempre referenciado com valores de outros ativos semelhantes.
- **Princípio da temporalidade ou de mercado:** O valor de uma propriedade é variável ao longo do tempo e as expectativas de rendimentos que poderá gerar no futuro são algo a ter em conta.
- **Princípio do justo valor:** A finalidade e o tipo de imóvel com que se faz uma avaliação condiciona o método e as técnicas a seguir. O valor de um imóvel pode ser único, contudo a sua avaliação pode ter vários valores consoante a sua finalidade.

3.1.3 Classificação da propriedade imobiliária

A classificação de um imóvel, do ponto de vista técnico-administrativo, na atual legislação portuguesa é feita através de três documentos:

- Código Civil - Decreto-Lei 47.344/66, de 25/11;
- CIMI - Código do Imposto Municipal sobre Imóveis - Decreto-Lei 287/2003, de 12/11;
- Código de Expropriações - Decreto-Lei 438/91, de 9/11 e Lei 168/99 de 18/09.

Código Civil

O código civil classifica como *'coisas imóveis'* (Artº 204): [4]

1. São coisas imóveis:

- *Os prédios rústicos e urbanos;*
- *As águas;*
- *As árvores, os arbustos e as fontes naturais, enquanto estiverem ligadas ao solo;*
- *Os direitos inerentes aos imóveis mencionados nas alíneas anteriores;*
- *As partes integrantes dos prédios rústicos e urbanos.*

2. Entende-se por prédio rústico uma parte delimitada do solo e as construções nele existentes que não tenham autonomia económica, e por prédio urbano qualquer edifício incorporado no solo, com os terrenos que lhe sirvam de logradouro.

3. É parte integrante toda a coisa móvel ligada materialmente ao prédio com carácter de permanência.

CIMI - Código do Imposto Municipal sobre Imóveis

A autoridade tributária e aduaneira define *'prédio'* (Artº 2) como: [5]

"toda a fracção de território, abrangendo as águas, plantações, edifícios e construções de qualquer natureza nela incorporados ou assentes, com carácter de permanência, desde que faça parte do património de uma pessoa singular ou colectiva e, em circunstâncias normais, tenha valor económico, bem como as águas, plantações, edifícios ou construções, nas circunstâncias anteriores, dotados de autonomia económica em relação ao terreno onde se encontrem implantados, embora

situados numa fracção de território que constitua parte integrante de um património diverso ou não tenha natureza patrimonial."

Os prédios classificam-se em rústicos, urbanos e mistos (Artº 3, 4 e 5), havendo ainda uma subdivisão dos prédios urbanos segundo a espécie (Artº 6).

Este artigo divide os prédios urbanos em:

- Habitacionais;
- Comerciais, industriais ou para serviços;
- Terrenos para construção;
- Outros.

Código de Expropriações

Este código trata separadamente de expropriações de solos e de edifícios ou construções e não inclui qualquer definição específica de prédio, nem de bens móveis e de direitos inerentes.

No entanto, a definição da propriedade constitui, entre outros, um dos parâmetros que concorrem para apreciação do seu valor. As propriedades podem ser designadas como **'Rústico'** e **'Urbano'**, sendo que a primeira designação é predominantemente terreno/solo, e a segunda é edifício/construção. [6]

3.1.4 Valor e preço na avaliação

O valor de um imóvel é uma entidade complexa que pode ser encarada sobre várias perspetivas. Existem duas lógicas sobre a natureza do valor: uma dita "univalente"e outra "plurivalente".

A primeira afirma que o valor de um determinado bem é único num dado momento, qualquer que seja a finalidade da avaliação. Por exemplo, as instituições financeiras, para segurança dos créditos a conceder, podem fixar para um imóvel objeto de empréstimo um valor de garantia.

A segunda lógica afirma que o valor pode mudar em função do objetivo da avaliação. É, portanto, um conceito mutável cujo significado varia em função da finalidade da avaliação.

Ambas as abordagens possuem relação com o conceito de preço. Preço é definido como sendo a expressão monetária de um bem, ou seja, a quantia em dinheiro que uma determinada mercadoria pode ser vendida. Desta forma, mesmo que dois imóveis possuam valores de mercado diferentes, eles podem ser vendidos pelo mesmo preço.

3.1.5 Finalidade da avaliação

Os bens imobiliários podem ser avaliados segundo diferentes perspetivas, não sendo fixos nem permanentes, logo é totalmente natural surgirem diferentes perspetivas nas análises de valor. De seguida são apresentadas as perspetivas mais correntes segundo as quais os bens imobiliários são objeto de avaliação:

- **Âmbito da atividade creditícia:** destacam-se as avaliações no âmbito do crédito hipotecário para segurança dos créditos a conceder; as instituições de crédito avaliam o imóvel com vista a financiar a sua aquisição ou a fixar garantias reais para cobertura de operações comerciais.

- **Âmbito das expropriações por utilidade pública:** a entidade que expropria, avalia ou solicita a avaliação do bem a expropriar tendo em vista determinar o montante a atribuir ao expropriado; este por seu lado, também poderá fazer ou solicitar a avaliação da propriedade em apreço, com vista a determinar a compensação a que tem direito.
- **Âmbito fiscal:** destacam-se as avaliações que visam a determinação do valor patrimonial dos prédios classificados segundo o Código do Imposto Municipal sobre Imóveis.
- **Âmbito da atividade seguradora:** estas avaliações visam o estabelecimento de prémios de seguros e são realizadas pelas seguradoras com o fim de determinar o valor do risco que irão cobrir.
- **Âmbito do processo civil:** avaliação tanto no processo executivo comum como no processo de execução universal (falência e insolvência), com o fim de se fixar o valor mínimo que irá servir como base de licitação na venda executa de imóveis ou direitos imobiliários.
- **Âmbito das transações:** estas avaliações são realizadas com uma certa regularidade nos processos de compra e venda de bens imobiliários; são requeridas pelo comprador e/ou vendedor do imóvel.
- **Âmbito dos fundos de investimento imobiliário:** o decreto-lei 252/2003, de 17 de outubro e decreto-lei 13/2005 de 07 de janeiro, concede uma importância acrescida às avaliações dos imóveis dos fundos de investimento imobiliário.

3.1.6 Tipos de valor

Para cada uma das perspetivas de avaliação e para o mesmo bem imobiliário corresponderá um valor não necessariamente igual ao das outras perspetivas. Indicam-se a seguir os tipos de valor mais utilizados na prática da engenharia de avaliação imobiliária.

- **Valor venal ou de capital:** é o valor em mercado livre pelo qual o bem foi transacionado (sendo assim referente ao preço do momento da transação).
- **Valor de mercado:** é o montante pelo qual se estima que uma propriedade adequadamente publicitada seja transacionada à data da avaliação entre um comprador e um vendedor interessados, cada um dos quais atuando independentemente do outro, sem coação e com conhecimento do mercado; a independência entre as partes envolvidas significa que não existem relações particulares ou familiares entre as mesmas que possam estabelecer um nível de preço não característico no mercado; a sua medida pode ser estimada com base no valor pelo qual se tem vindo a transacionar a maioria dos bens com características semelhantes às do bem em apreço.
- **Valor intrínseco:** definido como o custo necessário para a construção de um bem semelhante ou igual ao em apreço (incluindo custos com estudos, projetos, construção, taxas e demais encargos); esta medida é muito utilizada nas avaliações de propriedades nunca ou raramente transacionadas (hospitais, escolas, igrejas, etc.) e para as quais não existe valor de mercado.
- **Valor locativo ou de rendimento:** é o valor que resulta da capitalização a uma taxa conveniente dos rendimentos líquidos médios proporcionados pela propriedade, através de uma renda que é paga periodicamente.
- **Valor patrimonial:** é o valor atribuído pela autoridade tributária ao imóvel seja ele urbano ou rústico, e que consta na caderneta predial do prédio.

- **Valor residual:** é o valor de sucata ou de demolição que resta desse bem ao fim da sua vida útil.

3.1.7 Valor patrimonial tributário

O Valor Patrimonial Tributário expressa o valor real de um imóvel num dado ano. Encontra-se exposto no Código do Imposto Municipal sobre Imóveis (CIMI) e o consumidor pode consultar qual é o valor que se aplica à sua habitação na Caderneta Predial. O VPT usa-se tanto para calcular o IMI como o Imposto Municipal sobre as Transmissões Onerosas de Imóveis (IMT). [7]

É calculado da seguinte maneira:

$$\mathbf{VPT = Vc \times A \times Ca \times Cl \times Cq \times Cv}$$

Em que:

- **Vc:** Valor base dos prédios edificados, valor este que de acordo com o código 39º do CIMI é referente ao custo médio de construção por metro quadrado.
- **A = (Aa+Ab)*Caj+Ac+Ad:** Área bruta de construção (Aa+Ab) mais a área excedente à área de implantação (Ac+Ad).
- **Aa:** A área bruta privativa (Aa) é a superfície total medida pelo perímetro exterior e eixos das paredes ou outros elementos separadores do edifício ou da fracção, incluindo varandas privativas fechadas, caves e sótãos privativos com utilização idêntica à do edifício ou da fracção, a que se aplica o coeficiente 1.
- **Ab:** As áreas brutas dependentes (Ab) são as áreas cobertas e fechadas de uso exclusivo, ainda que constituam partes comuns, mesmo que situadas no exterior do edifício ou da fracção, cujas utilizações são acessórias relativamente ao uso a que se destina o edifício ou fracção, considerando-se, para esse efeito, locais acessórios as garagens, os parqueamentos, as arrecadações, as instalações para animais, os sótãos ou caves acessíveis e as varandas, desde que não integrados na área bruta privativa, e outros locais privativos de função distinta das anteriores, a que se aplica o coeficiente 0,30.
- **Ac e Ad:** A área do terreno livre do edifício ou da fracção ou a sua quota-parte resulta da diferença entre a área total do terreno e a área de implantação da construção ou construções e integra jardins, parques, campos de jogos, piscinas, quintais e outros logradouros, aplicando-se-lhe, até ao limite de duas vezes a área de implantação (Ac), o coeficiente de 0,025 e na área excedente ao limite de duas vezes a área de implantação (Ad) o de 0,005.
- **Ca:** O coeficiente de afetação varia entre 0,08 e 1,20 consoante o imóvel se destine a estacionamento, armazéns, indústria, habitação, comércio ou serviços – os valores a aplicar encontram-se pré-estabelecidos no artigo 41º do CIMI. [8]
- **Cl:** Por sua vez, o coeficiente de localização varia entre 0,4 e 3,5 (reduzindo-se o para 0,35 em meios rurais e nos casos de habitações dispersas). Para o cálculo deste coeficiente entram quatro grandes fatores de influência: existência de transportes públicos próximos da casa; acessibilidades, o que diz respeito à qualidade e variedade das vias rodoviárias, ferroviárias, fluviais e marítimas; presença de equipamentos e infraestruturas sociais, tais como escolas, comércio e serviços públicos; localização numa zona de elevado valor de mercado imobiliário.

- **Cq:** O coeficiente de qualidade e conforto – que oscila entre 0,5 e 1,7 – é aplicado ao valor base do prédio edificado, podendo ser majorado até 1,7 e minorado até 0,5, e obtém-se adicionando à unidade os coeficientes majorativos e subtraindo os minorativos que constam das tabelas do artigo 43º do CIMI. [8]

- **Cv:** O coeficiente de vetustez (Cv) é função do número inteiro de anos decorridos desde a data de emissão da licença de utilização, quando exista, ou da data da conclusão das obras de edificação, de acordo com a tabela no artigo 44º do CIMI. [8]

3.1.8 Preço de habitação e zonamento

O preço da habitação por metro quadrado de área útil em 2014, a formula de cálculo do preço de venda dos terrenos para habitação e as diferentes condições de alienação encontram-se na Portaria n.º 156/2014. [9]

Os preços da habitação por metro quadrado de área útil variam consoante as zonas do país:

- Na zona I - (euro) 679,35;
- Na zona II - (euro) 602,92;
- Na zona III - (euro) 557,91.

A Zona I contêm sedes de distrito e municípios das Regiões Autónomas, bem como Almada, Amadora, Barreiro, Cascais, Gondomar, Loures, Maia, Matosinhos, Moita, Montijo, Odivelas, Oeiras, Póvoa do Varzim, Seixal, Sintra, Valongo, Vila do Conde, Vila Franca de Xira e Vila Nova de Gaia.

A Zona II contêm Abrantes, Albufeira, Alenquer, Caldas da Rainha, Chaves, Covilhã, Elvas, Entroncamento, Espinho, Estremoz, Figueira da Foz, Guimarães, Ílhavo, Lagos, Loulé, Olhão, Palmela, Peniche, Peso da Régua, Portimão, Santiago do Cacém, São João da Madeira, Sesimbra, Silves, Sines, Tomar, Torres Novas, Torres Vedras, Vila Real de Santo António e Vizela.

A Zona III contêm os restantes municípios do continente.

O preço de venda dos terrenos destinados a programas de habitação de custos controlados é calculado pela aplicação da fórmula seguinte:

$$\mathbf{Pv = p \times Cf \times Au \times Pc}$$

Em que:

- **p:** variável entre 0,07 e 0,15, por forma diretamente proporcional à percentagem de infraestruturas executadas;
- **Cf:** fator relativo ao nível de conforto do fogo, conforme definido no artigo 2.º do Decreto-Lei n.º 329-A/2000, de 22 de dezembro, o qual é fixado livremente para as áreas não habitacionais não incluídas nos fogos;
- **Au:** área útil, determinada nos termos do Regulamento Geral das Edificações Urbanas (RGEU), quer para a parte habitacional, quer para a não habitacional, excluindo a área das garagens quando estas estejam incluídas nos fogos;
- **Pc:** (euro) 791,76 por metro quadrado de área útil para vigorar em 2014.

3.1.9 Mercado imobiliário

A avaliação de propriedades imobiliárias para estimativa quer de valores de troca quer de valores de uso, requer geralmente a necessidade de recolha de dados no mercado imobiliário apropriado.

Geralmente um mercado imobiliário é localizado (quando a propriedade é imóvel e condicionada pelo seu enquadramento físico) e segmentado (quando oferece diferentes recursos para as diferences necessidades dos utilizadores).

A segmentação pode ocorrer de vários formas, sendo que existem vários fatores que influenciam a avaliação de um imóvel, entre as quais estão:

- A localização da casa, incluindo o piso, a vista que é possível ter a partir da habitação, a orientação solar e ainda as acessibilidades;
- A qualidade da construção do imóvel;
- A data de construção;
- O estado de conservação (para casas usadas);
- O terreno onde está inserida;
- A tipologia e a disposição da habitação;
- Os acabamentos e equipamentos disponíveis;
- As facilidades da habitação, isto é, estacionamento, piscina, elevador, espaços verdes, entre outros;
- O estado do mercado, ou seja, a procura e a oferta atual.

No entanto, quando temos de decidir sobre qual método usar devemos refletir sobre vários fatores, nomeadamente:

- Tipo de direito ou interesse sobre a propriedade;
- A óptica ou perspectiva considerada para a avaliação;
- Nível de precisão exigido para a análise em questão;
- Número de transacções efectuadas e análogas ao imóvel em apreço;
- A natureza ou tipo de propriedade a avaliar, por exemplo, uma propriedade produtiva ou um imóvel para reabilitar.

3.2 Métodos para cálculo do valor imobiliário

Em termos de métodos mais convencionais para o cálculo de valores de imóveis deparamo-nos com 4 deles que são maioritariamente usados na atualidade. Estes vão ser explicados, resumidamente, em seguida, e são os seguintes:

- Método Comparativo
- Método do Rendimento
- Método do Custo

- Método do Valor Residual

Como o mercado imobiliário tem altos e baixos e é afetado por vários outros aspetos que não são particularmente relevantes, neste relatório, existem estes métodos já referidos em cima para calcular o valor de um tal imóvel, caso contrário, todos eles resultariam no mesmo valor.

3.2.1 Método Comparativo

O Método Comparativo (MC), também por vezes designado de método direto ou de comparação, fundamenta-se sobretudo no conhecimento do mercado local e dos valores pelos quais se têm vindo a transaccionar as propriedades com características semelhantes à que se pretende avaliar. A utilização deste método supõe 3 premissas:

1. A existência de um mercado imobiliário ativo;
2. A obtenção de informação correta;
3. A existência de transacções de imóveis semelhantes.

As propriedades podem ser semelhantes mas cada uma é única pois têm pormenores que as distinguem como localização, área bruta e envolvente, o estado da mesma, o fim a que se destina e o momento em que é feita a avaliação. A aplicação do método comparativo na sua forma mais evoluída utiliza as técnicas de homogeneização e de análise estatística.

As técnicas de homogeneização, ao ajustarem os dados recolhidos, permitem comparar propriedades que entre si apresentam caraterísticas diversas, nomeadamente em relação à idade, estado de conservação, área, localização geográfica, data de transacção e nível de acabamentos, uma vez que homogeneíza os dados recolhidos. Trata-se, em resumo, de recolher um número significativo de amostras no mercado, proceder ao seu tratamento de modo a poderem ser comparáveis (homogeneização de áreas, formas de pagamento, localizações, níveis de qualidade, e idade) [2].

As técnicas de análise estatística permitem, por sua vez, descrever a população dos dados, através da determinação de certos parâmetros estatísticos (média, moda, desvio padrão, percentis e outros), e possibilitam, para uma dada margem de segurança (confiança) enunciar um valor ou um leque de valores mais prováveis para o imóvel em apreço [2].

É de realçar que este método é bastante dependente das referências que se adquirem, uma vez que o método pretende justamente fazer a comparação de imóveis. Daí ser muito importante que haja muita informação e de qualidade para que o resultado seja de confiança. Quando não existem registos suficientes, pode levar a juízos de valor por haver um número restrito de transações, uma vez que, todos sabemos que existem negócios que são influenciados por razões pessoais ou até estratégicas. Por outro lado, a informação deve ser de qualidade pois, logicamente, caso contrário, podem levar a valores distorcidos na avaliação a realizar. A aplicação do Método Comparativo requer a sua execução em seis etapas:

1. Estabelecer um quadro de caraterísticas e qualidades próprias do imóvel a avaliar. No caso de edifícios de caráter histórico ou artístico, deverão ainda tomar em linha de conta o valor particular dos elementos que lhe conferem esse caráter;
2. Descrever o imóvel a avaliar com base nas características identificadas no passo 1;
3. Recolha de informação do segmento de mercado relativo aquele que se está a avaliar, identificando exemplos comparáveis ao objeto da avaliação por via da localização, uso e tipologia, baseando-se em informações concretas sobre transações reais, ou em ofertas de venda (um indicador não tão seguro), que tenham ocorrido num curto período de tempo anterior;

4. Seleção de uma amostra representativa de imóveis comparáveis à que se está a avaliar, rejeitando os casos de distorção produzidos por preços anormais ou por dados não comparáveis. Descrever os imóveis da amostra com base nas caraterísticas identificadas no passo 1;

5. Homogeneização dos preços unitários obtidos na amostra com o valor do imóvel a avaliar, considerando o fator tempo e as diferenças ou semelhanças das caraterísticas dos mesmos (área, idade, tipologia, envolvente, acabamentos, etc.);

6. Análise estatística dos preços unitários homogeneizados. Descrever os preços unitários da amostra em termos de parâmetros estatísticos;

7. Atribuição do valor da propriedade, em função dos preços de já homogeneizados, deduzindo todos os valores que não tenham sido tomados em conta nos passos precedentes.

3.2.2 Método do Rendimento

O Método do Rendimento (MR) também conhecido pelo "Método da Capitalização do Rendimento", é especialmente adequado para estimar os valores de propriedades produtivas ou que podem ser arrendadas (anual, mensal, semestral, sazonal) mediante um determinado valor de renda. O método do rendimento, é tal como o anterior, muito utilizado na estimação dos valores da propriedade imobiliária, sendo especialmente adequado nas seguintes situações:

- Avaliações de propriedades produtivas ou que podem ser arrendadas a determinado valor de renda (isto é, das quais se espera que forneçam um rendimento, em regra periódico- mensal, sazonal ou anual);
- Prédios urbanos (habitações, escritórios, unidades comerciais, etc.);
- Prédios rústicos (vinhas, pomares, eucaliptais, pinhais, etc.);
- Fixação de valores de trespasse em arrendamentos comerciais;
- Determinação do valor de utilização em direito de superfície;

A avaliação de um bem imobiliário com recurso a esta metodologia engloba basicamente os seguintes passos:

1. Estimativa dos rendimentos brutos médios esperados;
2. Estimativa dos rendimentos líquidos médios proporcionados pela propriedade em apreço;
3. Fixação da taxa de actualização ou de capitalização (conforme a abordagem escolhida);
4. Cálculo do valor da propriedade em avaliação.

O valor resultante deste processo depende da estimativa dos rendimentos e do grau de adequação da taxa fixada. Apesar do rendimento e das taxas serem igualmente importantes para determinar o valor do imóvel através deste método a segunda oferece mais dificuldades pois é onde há menos consenso, ao definir e aceitar valores embora a outra também dê algumas dores de cabeça. O MR estima o valor atual dos rendimentos futuros derivados da propriedade, utilizando-se o princípio da antecipação. No âmbito do MR utilizam-se normalmente dois métodos de capitalização, o de capitalização direta (CD) e o *cash-flow* descontado (CFD). O Método da Capitalização Direta é utilizado para converter a estimativa da renda de um único ano numa indicação de valor através de uma operação direta dividindo o rendimento anual líquido por uma taxa de rendimento

ou taxa de capitalização apropriada que parte do pressuposto que a renda será perpétua e, em termos anuais, constante. Já o método do *cash-flow* descontado, parte do pressuposto que o projeto será desenvolvido e alugado ou vendido durante um certo período de tempo considerado razoável para a sua conclusão e absorção no mercado, sendo o valor final do imóvel traduzido pelo valor actual dos benefícios futuros líquidos inerentes (receitas - custos). Para os imóveis já existentes, o raciocínio é equivalente, baseando-se nas rendas líquidas que estão a ser praticadas ou, estando desocupado, nas rendas que potencialmente poderão ser praticadas e no valor de revenda após o período de investimento considerado. Os custos de desenvolvimento ou adaptação e outros encargos são subtraídos aos resultados estimados da venda ou do aluguer. A projeção do montante assim calculado, o rendimento líquido é descontado no período de tempo considerado (período de estudo), a uma taxa de desconto ou de actualização que reflecte a rentabilidade esperada e o risco inerente à propriedade considerada.

Os determinantes do valor no MR incluem um conjunto de factores dos quais destacamos os seguintes:

- A renda paga e o valor da renda;
- A taxa de capitalização apropriada para a propriedade em apreço;
- Os custos de gestão, reparação, seguros, taxas;
- Os termos do arrendamento;
- As garantias dadas pelos arrendatários;

Os rendimentos podem ser brutos ou líquidos, sendo que, o primeiro é igual ao valor anual ou (do período de tempo considerado) da renda contratual (caso dos prédios arrendados), ou ao resultante da multiplicação da produção anual pelo preço de mercado dos produtos agrícolas ou florestais produzidos (caso das parcelas produtivas agrícolas ou florestais em prédios rústicos), ou ainda (caso dos prédios urbanos não arrendados), ao valor anual da renda de propriedades arrendadas análogas e que possam servir de referência. Por outro lado, os rendimentos líquidos verificados no período de tempo considerado obtêm-se subtraindo esse rendimento bruto a todas as despesas e encargos que tiverem lugar nessa unidade de tempo. Para obtermos a taxa de capitalização podemos dividir o rendimento (bruto ou líquido) no período de tempo e o valor de transação de que se verificam em propriedades análogas a que está a ser sujeita a avaliação, obtendo assim a taxa de capitalização (bruta ou líquida, dependendo do tipo de rendimento). Para o calculo da propriedade em questão deve fazer-se a divisão do rendimento líquido ou bruto pela taxa de capitalização líquida ou bruta, respetivamente, no caso da capitalização direta.

O valor de um bem imóvel é obtido através da actualização dos rendimentos, efectivos ou previsíveis, que o mesmo gera ou é susceptível de gerar no futuro, ou seja:

$$VA = \frac{R_1}{(1+t_a)} + \frac{R_2}{(1+t_a)^2} + \frac{R_3}{(1+t_a)^3} + ... + \frac{R_n}{(1+t_a)^n} + \frac{VR}{(1+t_a)^n}$$

Ou

$$VA = \sum_{i=1}^{n} \frac{R_i}{(1+t_a)} + \frac{VR}{(1+t_a)^n}$$

onde:

VA = Valor actual do imóvel,
R_i = Rendimento líquido no ano t ; (CFs)
VR = Valor residual
t_a = taxa de actualização anual nominal
n = número de períodos de tempo

Figura 3: Fórmula para cálculo de um imóvel segundo o método do rendimento [2]

Deve-se estimar os rendimento líquido tendo em conta que o valor do imóvel a calcular deve ser o mesmo depois dos impostos.

A dificuldade do método de rendimento tradicional assenta precisamente na fixação dos valores do rendimento líquido (RL) e da taxa de capitalização (tc). O primeiro pode ser obtido, se a propriedade estiver arrendada a partir do rendimento que é obtido no momento atual ou, no caso da propriedade estar vaga, a partir de uma estimativa do valor da renda obtida através das rendas pagas por propriedades comparáveis à propriedade em apreço. O segundo, a taxa de capitalização, é obtida através da análise das vendas de propriedades comparáveis à propriedade em apreço. Em consequência disso, um avaliador deverá ter um conhecimento dos dois mercados distintos onde a propriedade em questão é colocada: o mercado de arrendamento e o mercado das vendas. Ou seja, tal como no método anterior, a quantidade e qualidade da informação do mercado imobiliário é fundamental para a correta adequação deste método.

3.2.3 Método do Custo

O Método do Custo (MdoC), assenta sobretudo na estimação do custo de reprodução ou de substituição da propriedade em apreço. O valor do imóvel, na ótica deste método, obtém-se adicionando ao valor de mercado do terreno (obtido com base na utilização do "Método Comparativo") e respectivos encargos com a sua aquisição, o custo da construção eventualmente depreciado em função da obsolescência física e/ou funcional e/ou ambiental e/ou económica detectadas, e/ou apreciado em função de singularidades arquitectónicas, históricas, ou outras verificadas. Este método tem particular interesse nas seguintes situações:

- Avaliação de propriedades nunca ou raramente transaccionadas e não vocacionadas para o lucro - hospitais, edifícios escolares, edifícios prisionais, bibliotecas, museus, castelos, etc;
- Avaliação de edifícios antigos;
- Avaliação de construções ou partes de construções para efeitos de fixação de prémios de seguro, indemnizações e tributações fiscais e outros;

- Avaliação de obras para reabilitação.

Podemos concluir então que, resumidamente, este método é usado em grande parte das situações onde a propriedade em causa nunca ou raramente muda de mãos. Portanto uma análise de preço de transações não será útil pois não existem registos que a suportem. O "custo de reprodução"corresponde ao custo da realização de uma obra idêntica à que se está a avaliar enquanto que o "custo de substituição"é relativo ao custo de realização de uma obra análoga. Por razões óbvias, o custo de substituição é particularmente útil para estimar o valor de propriedades onde é impensável a ideia de uma construção utilizando os mesmos materiais e processos utilizados na tal construção (exemplos disto são mosteiros, castelos, catedrais, entre outras. A aplicação do MdoC, quer na determinação do valor das construções, quer de terrenos com potencialidade construtiva, urbanos ou por lotear, passa pela definição de todos parâmetros que conduzem ao valor final, nomeadamente:

- Custos de demolição ou reabilitação de construções eventualmente existentes que se tornem necessárias eliminar ou/e reconstruir;
- Custos de construção dos imóveis e ou das infra-estruturas necessárias;
- Custos de estudos, projectos, licenças, fiscalização, taxas, assistência técnica;
- Custo de marketing, comercialização e vendas;
- Custos de financiamento;
- Valor da depreciação física e funcional, quando se trata de construções não novas;
- Margens de lucro exigíveis.

Este método não tem em conta os mecanismos da oferta e da procura (sinónimo de variação dos valores dos bens) pelo que este apresenta a vantagem de separar o valor de mercado do valor económico real da propriedade. A aplicação deste método resume-se nos seguintes passos:

1. Determinação do valor do terreno (Método Comparativo);
2. Estimativa do Custo Global de Construção (de substituição ou de reprodução);
3. Cálculo do Custo Global de construção depreciado/apreciado;
4. Cálculo do encargos de comercialização e lucro;
5. Cálculo do Valor do Imóvel.

Com base nas considerações anteriores, poderemos fixar o seguinte formulário como suporte da aplicação do "Método do Custo":

$$V = (T + E_T) + (C + E_C) \times (1 - K_{FI}) \times (1 - K_{FU}) \times (1 - K_{FA}) \times (1 + A) + E_{COM} + L$$

sendo :

V	- valor do empreendimento no estado em que se encontra;
T	- valor de mercado do terreno já infra-estruturado;
E_T	- encargos conexos com a aquisição do terreno;
C	- preço da construção (reprodução ou substituição);
Ec	- encargos conexos com a construção;
K_{FI}	- coeficiente de depreciação física;
K_{FU}	- coeficiente de depreciação funcional;
K_{AM}	- coeficiente de depreciação ambiental;
A	- coeficiente de apreciação;
E_{COM}	- encargos com a comercialização;
L	- lucro do promotor.

Figura 4: Fórmula para cálculo de um imóvel segundo o método do custo [2]

3.2.4 Método do Valor Residual

O "Método do Valor Residual"é um caso particular do "Método do Custo"pois considera no seu processo de cálculo todos os custos e receitas - envolvidos na execução do empreendimento imobiliário. Este método aplica-se na estimativa do valor de bens imobiliários com um valor potencial, ou seja, cujo valor poderá ser substancialmente superior se forem investidos capitais de modo a promover a sua alteração ou ampliação. Nesse sentido, este método aplica-se correntemente na estimação do valor de [2]:

- propriedades que irão ser objecto de obras de beneficiação (alteração, ampliação ou outras) e de que se pretende conhecer o valor, no estado físico em que se encontram;
- terrenos.

Tendo em conta estes aspetos podemos dizer que é útil saber quais os proveitos potênciais do imóvel uma vez que este método distingue-se por avaliar obras que irão ser sujeitas a obras para poderem ser melhoradas ou ampliadas. O método funciona na base da premissa que o preço do qual um comprador pode pagar por tal propriedade é o excedente que resulta depois de ele ter deduzido da venda do desenvolvimento (empreendimento), acabado os custos do desenvolvimento, incluindo os custos de projecto e construção, os custos de aquisição e venda, encargos financeiros, taxas e lucro requerido para levar a cabo o projecto. O método pode ser expresso como se segue: Produto da venda – Menos custos de desenvolvimento e lucro = Excedente para terreno. Portanto o MR procura determinar o excedente disponível depois de deduzidos todos os custos de desenvolvimento e o lucro do promotor.

O cálculo do valor das propriedades com recurso a este método pode seguir a abordagem estática ou dinâmica sendo que na primeira não se considera a inflação nem a atualização dos valores dos "cash flows"enquanto que a segunda utiliza essa mesma atualização com base numa taxa de atualização adequada (já falada no método do rendimento). Seguindo a abordagem estática temos os seguintes passos:

1. Estimativa do valor comercial do imóvel depois da intervenção de reabilitação utilizando o Método do Rendimento - Capitalização Directa (PC);
2. Estimativa dos custos associados à intervenção e fixação do valor do Lucro (C e L);
3. Cálculo do valor do imóvel no estado actual (sem a intervenção) V = PC - (C + L).

Por outro lado, utilizando a abordagem dinâmica temos:

1. Estimativa das receitas esperadas do imóvel depois da intervenção de reabilitação ou redesenvolvimento utilizando a técnica do cash-flow descontado (Ri);
2. Estimativa dos custos associados à intervenção e fixação do valor do Lucro utilizando a técnica do cash-flow descontado (Ci e L);
3. Cálculo do valor actual líquido do imóvel no estado actual.

$$VA = \sum_{t=0}^{N} \frac{Ri - C}{(1+d)^N} - L$$

Relativamente ao cálculo dos valores de terrenos, podemos fazer a diferença entre o valor comercial presumível do empreendimento acabado e o somatório de todos os custos que o empreendimento imobiliário suportou, ou que se prevê vir a suportar, ao longo de todo o desenvolvimento do empreendimento (com exclusão, evidentemente, do custo de aquisição do próprio terreno).

3.3 Utilizações de modelos de *Machine Learning* na previsão de valor imobiliário

Os modelos de previsão de valor imobiliário não são algo muito recente no panorama do Machine Learning. De facto, já foram efectuados vários projetos utilizando algoritmos como redes neuronais artificais, regressão hedónica, AdaBoost e a árvore J48, que são considerados os melhores modelos base para previsão de preços. Estes modelos, juntamente com ferramentas de data mining, conseguiram sempre gerar resultados com uma boa precisão.

Em 2010, Hu Xiaolong e Zhong Ming falaram sobre a necessidade do desenvolvimento de projetos para a previsão de valores imobiliários, referindo, já na altura, que os recursos computacionais e as ferramentas disponíveis conseguiriam resolver o problema que as imobiliárias enfrentam atualmente: a imprevisibilidade do mercado e a enorme quantidade de variáveis a ter em conta.

Já em 2016, Yang Li para a *35th Chinese Control Conference (CCC)*, discursou sobre um projeto para o conhecido site de arrendamento de imóveis "AirBnb"através da utilização de um modelo de regressão linear. Segundo ele, um dos factores importantes nesta previsão e no caso destes imóveis era o agrupamento dos dados segundo a distância a que se encontram de atrações turísticas. Contudo, é preciso ter em conta que a plataforma em questão é mais utilizada por turistas daí esta importância. Assim, de uma forma geral, percebeu-se que o agrupamento dos dados segundo a localização relativamente a pontos de interesse seria importante. [10]

Na *17th International Conference*, Li e Chu explicaram a sua tentativa de criação de um modelo de previsão que pudesse ajudar os bancos a providenciar os créditos para os clientes. Segundo os autores após serem recolhidos todos os dados necessários era utilizado um algoritmo de redes neuronais para prever o preço. Para quantificar a precisão do algoritmo eram utilizados 2 valores: a raiz quadrada do erro-médio e o erro percentual absoluto médio. [11]

Não só nos países asiáticos é que se verificou este interesse pela previsão dos valores imobiliários. Durante 2005, devido ao grande aumento do interesse pelos mercados imobiliários nos EUA, foi atingido um ponto de falência nestes mercados. Com isto os valores dos imóveis tiveram descidas muito grandes, entre os 30% a 60% nas grandes cidades que se prolongou durante vários anos. [12]

Com cada vez menos investimento a ser feito nestes mercados, a partir de Novembro de 2012, estes começaram a recuperar o que levou Park e Bae, em 2015 a pesquisarem e tentarem criar um modelo de previsão que lhes permitisse prever o que os imóveis iriam valer no futuro. Neste modelo foram utilizados vários algoritmos de Machine Learning: C4.5, RIPPER que seleciona uma classe majorante e uma classe minorante, Naive Bayesian que divide o dataset em diferentes classes através da distribuição da probabilidade, AdaBoost que é um algoritmo que tem como objetivo aumentar a qualidade dos outros algoritmos e é feita a comparação entre todos estes chegando-se à conclusão que entre estes, em termos de previsão, o RIPPER é o que melhores resultados apresenta.

4 Estado da arte

Para um bom desenvolvimento de qualquer projeto é necessário possuir, inicialmente, um bom conhecimento do desenvolvimento tecnológico existente até ao momento, assim como o que está em produção. O estado da arte é, portanto, um fase fulcral de qualquer projeto pois providencia uma base de conhecimento, neste caso, sobre previsão de valor imobiliário.

O mercado imobiliário é um dos mais competitivos em termos de preços e o mesmo tende a variar significativamente com base em vários fatores. Desta forma, a aplicação dos conceitos de *machine learning* e de outras técnicas para otimizar e prever os preços com alta precisão (*target*) torna-se um dos campos principais na área.

4.1 Modelos de regressão

Nesta secção são apresentados vários fatores importantes que devem ser utilizados para prever os preços das casas com boa precisão. Neste casos, os modelos de regressão utilizam várias *features*, e tentam encontrar quais provocam um menor erro na soma dos quadrados residuais. Ao utilizar *features* num modelo de regressão, é necessária alguma engenharia para uma melhor previsão. Frequentemente, um conjunto de *features* (regressões múltiplas) ou regressão polinomial (aplicando vários métodos nas *features*) é utilizado para ajustar o modelo da melhor forma. Numa regressão linear simples, é utilizada apenas uma *feature* para treinar o modelo e tentar prever o preço final da casa. [13]

A equação de ajuste do modelo é linear, e portanto, definida como:

$$Preco(feature) = a + w1 \times feature$$

Nos modelos de regressão múltipla, em vez de uma só *feature*, são utilizadas várias e variadas, no sentido de encontrar as que melhor se conjungam no sentido de minimizar o erro na soma dos quadrados residuais.

$$Preco(feature) = a + w_1 \times feature_1 + w_2 \times feature_2 + ... + w_n \times feature_n$$

Os modelos de regressão múltipla normalmente são utilizados para cenários de previsão. Estes modelos podem ser utilizados para explicar mudanças em variáveis independentes e dependentes, bem como para avaliar a importância relativa de cada variável independente. Cada coeficiente de regressão estima a quantidade de mudança que ocorre na variável dependente para uma mudança de unidade nas variáveis independentes na equação. Nesta área tal é feito da seguinte forma:

- Concetualização do modelo: A consideração da teoria relevante, que ajudará a determinar os fatores relevantes ou variáveis independentes para explicar a variável dependente no modelo.
- Estimativa e desenvolvimento do modelo: o modelo pode ser estimado por testes estatísticos para garantir que o modelo seja rigoroso.
- Teste de modelo: isso envolverá a aplicação de testes estatísticos relevantes. A qualidade do modelo é avaliada usando o coeficiente de determinação, que é indicado como R2. Quanto maior o R2, melhor o modelo.
- Previsão/explicação: O modelo pode então ser testado e validado na outra amostra.

Podem ser definidos vários modelos com variadas *features* e variadas complexidades aplicadas aos modelos, no entanto a utilização de poucas *features* leva a um *underfit*, enquanto que um modelo com bastante maior complexidade leva facilmente a um *overfit*. O mais normal tem sido realizar o segundo método, e corrigir o *overfit* com outras técnicas como o *LASSO*. A previsão de preços de imóveis também tem utilizado a pesquisa do *K Nearest Neighbourgh*, que tende a agrupar vários imóveis do mesmo género, de forma a mais tarde agrupar imóveis do mesmo preço.

4.2 Modelos ARIMA

Em análise de séries temporais, um modelo auto-regressivo integrado de médias móveis (ARIMA, na sigla em inglês) é uma generalização de um modelo auto-regressivo de médias móveis (ARMA). Ambos os modelos são ajustados aos dados da série temporal para entender melhor os dados ou para prever pontos futuros na série. [14]

Combinando um modelo AR (processo *Auto Regressive*) com um modelo MA (processo *Moving Average*), um modelo ARIMA é construído. O valor de uma variável num modelo autoregressivo depende dos valores passados da variável mais uma percentagem de erro. Num AR, existe um parâmetro `p`, que denota o número de defasagens consideradas no modelo. O parâmetro ruído é indicado como u_t.

$$y_t = u + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + u_t$$

É uma propriedade desejável que um processo de AR(p) seja estacionário. Isso significa que a autocorrelação irá convergir para zero após um número limitado de defasagens. Se não for esse o caso, uma raiz unitária pode estar presente, o que é indesejável para fins de previsão.

O modelo de média móvel é uma regressão linear nos termos de erro atual e atrasado. Os termos de erro são assumidos como independentes entre si, para seguir uma distribuição normal, com média constante e variação constante.

$$y_t = u + u_1 + \theta_1 u_{t-1} + \theta_2 u_{t-2} + ... + \theta_q u_{t-q}$$

Um processo finito de MA é sempre estacionário, possui uma média constante, uma variação constante e a autocovariância pode ser diferente de zero até retardar `q`, e será zero a partir de então.

Um modelo de média móvel autoregressiva ARMA (p, q) é criado combinando um AR (p) com um modelo MA (q). Nesse modelo combinado, o valor atual é linear dependente dos seus próprios valores passados e dos valores atuais e passados do parâmetro no modelo de ruído branco.

$$y_t = u + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + ... + \theta_q u_{t-q} + u_t$$

O parâmetro ruído branco é normalmente distribuído, tem uma média de zero e variação constante. O I no ARIMA significa integrado. Como a variável subjacente ao modelo ARMA precisa ser estacionária, às vezes é necessário aplicar diferenciação para atender a essa procura. Um modelo ARMA (p, q) de uma variável d vezes diferenciada é, portanto, o mesmo que um modelo ARIMA (p, d, q).

4.3 Modelos de *Machine Learning*

Muitos algoritmos de *Machine Learning* são usados para aumentar efetivamente a precisão da previsão. Vários investigadores fizeram pesquisas e implementaram algoritmos como regressão hedônica, redes neuronais artificiais, *AdaBoost*, *J48 Tree*, sendo estes os modelos considerados como os melhores na previsão de preços. Estes são considerados os modelos base e com a ajuda de

algoritmos avançados de ferramentas de mineração de dados alcançam uma taxa maior de precisão da previsão. São exemplos desses algoritmos o *Random forest*, o *Gradient boosted trees*, as *Artificial neural networks* e os *Ensemble learning models*. Os resultados e a avaliação desses modelos usando *machine learning* e ferramentas avançadas de mineração de dados, como o *Weka* e o *Rapid Miner*, têm uma grande influência na previsão de preços. [15] [16]

4.3.1 Random Forest

O *Random Forest* pode ser utilizado tanto para prever a classificação como a regressão. O processo principal é o desenvolvimento de muitas árvores de decisão com base na seleção aleatória de dados e na seleção aleatória de variáveis, e fornecimento da classe de variável dependente baseada em muitas árvores. A principal vantagem de usar este algoritmo nos dados provêm do facto deste manipular os valores em falta, e ainda assim manter a precisão desses dados, ainda com grande probabilidade do *overfitting* ser baixo. Nas árvores de regressão, o resultado será contínuo.

4.3.2 Support Vector Machines

O *Support Vector Machines* também pode ser usada como uma técnica de regressão. Este produz um hiperplano que separa os pontos com diferentes *features*. Na *Linear Support Vector Regression* ele estima uma função que maximiza o desvio do *target* na faixa de margem normalizada, mantendo a função o mais plana possível. Portanto, a técnica de *Support Vector Regression* é um problema de minimização convexa que encontra o vetor normal da função linear.

4.3.3 Gradient Boosting

O *Gradient Boosting* pode ser utilizado tanto para regressão como para classificação. O aumento de gradiente é uma técnica para produzir modelos de regressão consistindo em coleções de regressores. É uma instanciação dessa ideia de regressão. O tema principal é seguir repetidamente o procedimento simples de previsão da regressão dos dados, computando o erro residual. Através da quantidade do erro por dados, é criado, com esse conhecimento, um novo modelo para prever o erro residual. O conceito principal é a criação dum conjunto de previsões, depois localização dos erros e redução destes mesmos.

4.3.4 Artificial Neural Networks

Nas *Artificial Neural Networks*, o MLP (*multi layer perceptron*) é um perceptron multicamada que possui a mesma estrutura de um percepton de uma única camada, só que com uma ou mais camadas ocultas. As camadas ocultas são conectadas diretamente à camada de entrada, na qual os valores de entrada são apresentados em perceptrons. Os perceptrons classificam qualquer conjunto de entradas se os valores de entrada forem apresentados ao perceptron. Se a saída prevista for igual à saída desejada, o desempenho será considerado satisfatório e nenhuma alteração nos pesos será feita. Se não corresponder, os pesos são alterados para reduzir o erro.

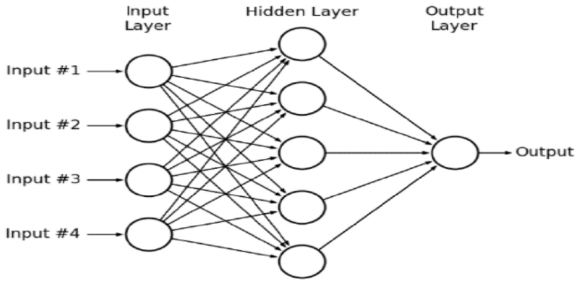


Figura 5: Modelo *multi layer perceptron*

4.3.5 Decision Trees

As *Decision Trees* são consideradas como um dos melhores e mais utilizados algoritmos de aprendizagem supervisionado. Este modelo tem a capacidade de prever a saída com imensa de precisão e estabilidade. É usado para prever qualquer tipo de problema, como classificação ou regressão. No entanto, neste caso, como estamos perante um *target* contínuo, o problema é do tipo regressão.

Nesse modelo, o conjunto de dados disponível pode ser contínuo ou categórico. É utilizada uma árvore binária que particiona recursivamente o vetor de previsão em diferentes subconjuntos, de modo que o *target* seja mais homogéneo. As restantes *features* representam o vetor dos preditores. Uma árvore de decisão com nós terminais é usada para comunicar a decisão de classificação. Um parâmetro = (1, 2, 3, ... t) associa o valor do parâmetro i (i = 1, 2, 3, ..., t) ao i-ésimo nó terminal. O procedimento de particionamento pesquisa todos os valores das variáveis de previsão (vetor de preditores) para encontrar a variável que fornece a melhor partição nos nós filhos. A melhor partição será a que minimizará a variação ponderada. No entanto, um dos principais desafios nas árvores de decisão é o excesso de ajustes. Na pior das hipóteses, ele considerará o nó da folha para cada valor e, assim, fornecerá 100% de precisão. Para evitar o ajuste excessivo, é possível definir restrições no tamanho da árvore ou corta-la.

4.3.6 Ensemble Learning Bagging

O *Ensemble Learning Bagging* é um algoritmo designado para melhorar a estabilidade e reduzir a variação. É uma aplicação de aprendizagem em grupo com construção de vários modelos, que se juntam e criam um modelo mais preciso. No *bagging* vários modelos são construídos em paralelo em várias amostras e, em seguida, os vários modelos votam para fornecer o modelo final e, consequentemente, a previsão.

5 Levantamento de requisitos do sistema

Para realizar a conceção de um sistema completo e com um bom funcionamento é fulcral, antes da implementação, construir um plano de requisitos referentes à aplicação e às funcionalidades que esta necessita. Após recolha de dados, o sistema possui duas principais funcionalidades: classificar um imóvel com um preço consoante determinadas *features*; prever a tendência futura do preço de um imóvel.

Deste modo, achamos por bem dividir os requisitos do sistema em três principais fases:

- Recolha de dados e informação;
- Identificação do preço:
  - Pré-processamento e análise dos dados;
  - Aplicação de técnicas e algoritmos de regressão;
  - Análise da regressão;
- Previsão da tendência futura do preço:
  - Aplicação de técnicas de previsão do *target*;
  - Análise preditiva.

Nas subsecções deste capítulo vão ser enumerados, portanto, os requisitos relacionados com cada fase.

Importante referir que, como o caso de estudo é o concelho de Braga, todos os dados estão já limitados a esta *feature*, neste caso, a localização.

Recolha de dados e informação

O sistema deve:

- Recolher informação de vários *websites* relacionados com o mercado imobiliário, utilizando práticas de *webscraping*;
- Abranger qualquer tipo de imóvel, nomeadamemnte:
  - Apartamentos;
  - Moradias;
  - Lojas;
  - Armazéns;
  - Entre outros.
- Abranger o número possível de *features* disponibilizadas pelos *websites*, nomeadamente:
  - Finalidade (venda ou arrendamento);
  - Número de quartos;
  - Número de casas de banho;
  - Número de pisos;
  - Presença de garagem;

  - Presença de piscina;
  - Área bruta;
  - Estado (novo ou usado);
  - Localização (freguesia);
  - Serviços na proximidade;
  - Preço de venda;
  - Entre outros.
- Colocar os dados recolhidos num *dataset*, de forma organizada por *features*.

Pré-processamento e análise dos dados

O sistema deve preparar os dados, transformando-os de forma a que as técnicas e os algoritmos utilizados de seguida na regressão sejam mais efetivos.

Desta forma, o sistema deve:

- Produzir gráficos representativos dos dados;
- Verificar a existência de *outliers* nos dados;
- Verificar a existência de valores em falta nos dados;
- Remover colunas com mais de x% de valores em falta;
- Remover linhas com mais de x% de valores em falta;
- Inserir dados nos campos com valores em falta, utilizando métodos estatísticos para cálculo dos novos dados;
- Realizar standardização dos dados, caso seja vantajoso;
- Realizar discretização dos dados, caso seja vantajoso;
- Realizar normalização dos dados, caso seja vantajoso;
- Realizar seleção de *features*, através de técnicas como:
  - Remoção de *features* com baixa variância;
  - Seleção univariada de *features*;
  - Eliminação recursiva de *features*.

Aplicação de técnicas e algoritmos de regressão

Após a aplicação de diferentes técnicas para pré-processar os dados, o sistema deve proceder à aplicação de diferentes combinações de técnicas e modelos.

O sistema deve, portanto:

- Dividir os dados, criando dois conjuntos, sendo um deles os dados de treino e outro os dados de teste;
- Retirar o *target* de ambos os conjuntos de dados, neste caso, o preço de venda do imóvel;

- Aplicar modelos de regressão aos dados de treino, nomeadamente:
  - Linear Regression;
  - Random Forest;
  - Ridge Regression;
  - Gradient Boosting;
  - Decision Trees with AdaBoost;
  - LightGBM;
  - Artificial Neural Networks.

Análise da regressão

Após aplicação dos modelos, o sistema deve:

- Aplicar os respetivos modelos de regressão, que foram utilizados nos treinos, aos dados de teste, de forma a avaliar a precisão de cada modelo.
- Avaliar, de entre os modelos, o que proporciona resultados com melhor precisão para a regressão, através dos resultados obtidos para o *target*, fazendo uso de métricas de avaliação de regressão.

Aplicação de técnicas de previsão do *target*

Para a previsão da tendência futura do preço, o sistema deve:

- Criar coleção com base no `id` e `preço` dos dados, após breve preparação dos mesmos;
- Produzir gráficos representativos dos preços em função do tempo;
- Aplicar técnicas de *Machine Learning* que permitam prever séries temporais.

Análise preditiva

Nesta fase final, o sistema deve:

- Avaliar os resultados obtidos da previsão efetuada para a série temporal.

6 Recolha de dados e informação

De forma a poder atingir o objetivo final do trabalho, isto é, classificar e prever o preço dum imóvel consoante determinadas *features*, foi necessário numa fase inicial realizar a recolha dos dados para garantir o funcionamento de todo o sistema, pois estes são o fator principal em qualquer projeto de *Machine Learning*. Neste capítulo em específico é descrito o processo da recolha de dados do seu estado mais bruto até à organização destes em ficheiros.

6.1 Fontes de informação

Para iniciar o nosso processo de recolha de dados e toda a informação relevante para o nosso projeto foi necessário, inicialmente, fazer uma pesquisa para encontrar todas as páginas às quais seria possível aceder e recolher a informação que precisávamos sem quaisquer tipo de problema. Decidimos procurar pelos sites imobiliários com maior reputação na área. Alguns dos sites sugeridos, infelizmente, possuíam barreiras que impossibilitavam aplicar um algoritmo de webscraping ao mesmo como, por exemplo, *captcha codes*. Assim sendo, desta pesquisa, resultaram 2 sites imobiliários: **ERA** e **Imovirtual**. Em cada um dos sites é possível verificar estruturas fixas com a informação disponibilizada, sendo essa apresentada de seguida.

6.1.1 ERA

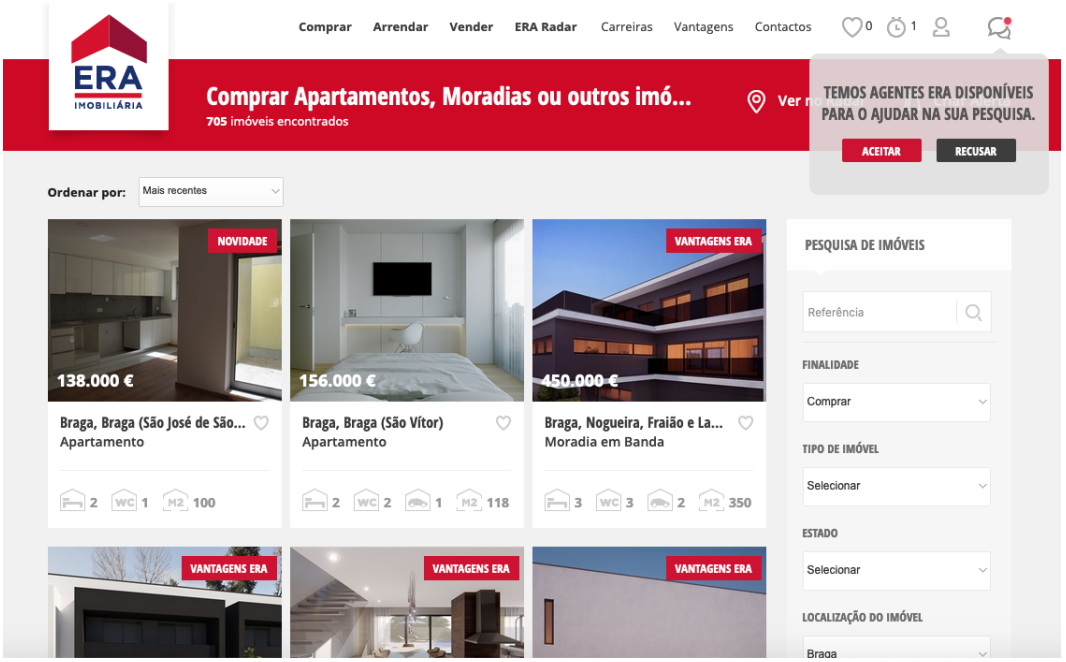


Figura 6: Visão geral dos imóveis do site da Era

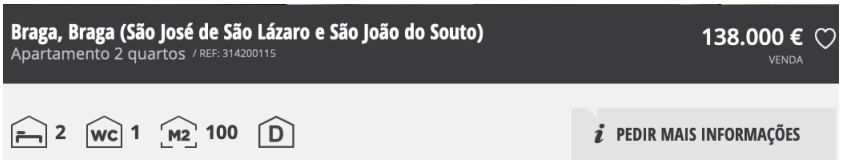


Figura 7: Características iniciais do site da Era

Finalidade:	**Venda**	Distrito:	**Braga**
Tipo de imóvel:	**Apartamento**	Concelho:	**Braga**
Estado:	**Usado**	Freguesia:	**Braga (São José de São Lázaro e São João do Souto)**
Preço de venda:	**138.000 €**	Ref:	**314200115**
Área bruta:	**100m²**		

Figura 8: Características principais do site da Era

Geral	Nº Frentes: 2; Nº Pisos: 4;
Água	Companhia;
Caixilharia	Dupla, Vidro Duplo; Estores: Eléctricos;
Domótica	Gás Canalizado;
Edifício	Tipo empreendimento: Habitação; Pavimento: Tijoleira; Paredes: Pintadas; Tecto: Estucado;
Exposição Solar	Nascente, Poente;
Vistas	Jardins, Rio;
Zona	Proximidade: Bancos, Centros Comerciais, Clínica, Escolas, Farmácia, Ginásio, Jardins, Jardins Infância, Padaria, Serviços Públicos, Supermercado;

Figura 9: Características adicionais do site da Era

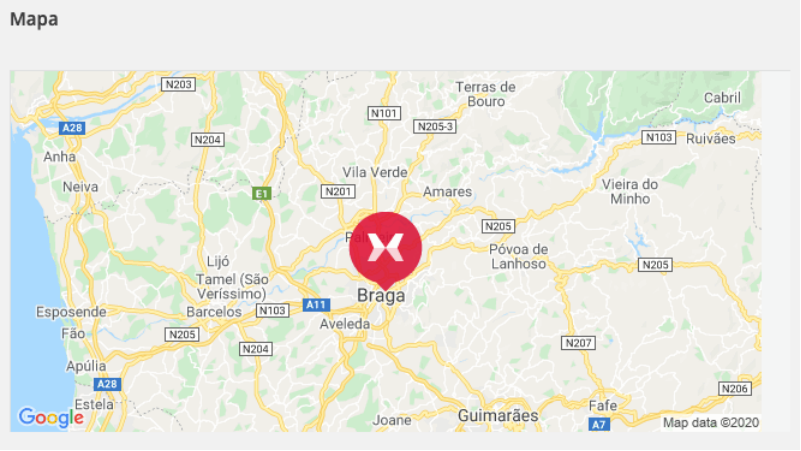


Figura 10: Coordenadas geográficas do site da Era

6.1.2 Imovirtual

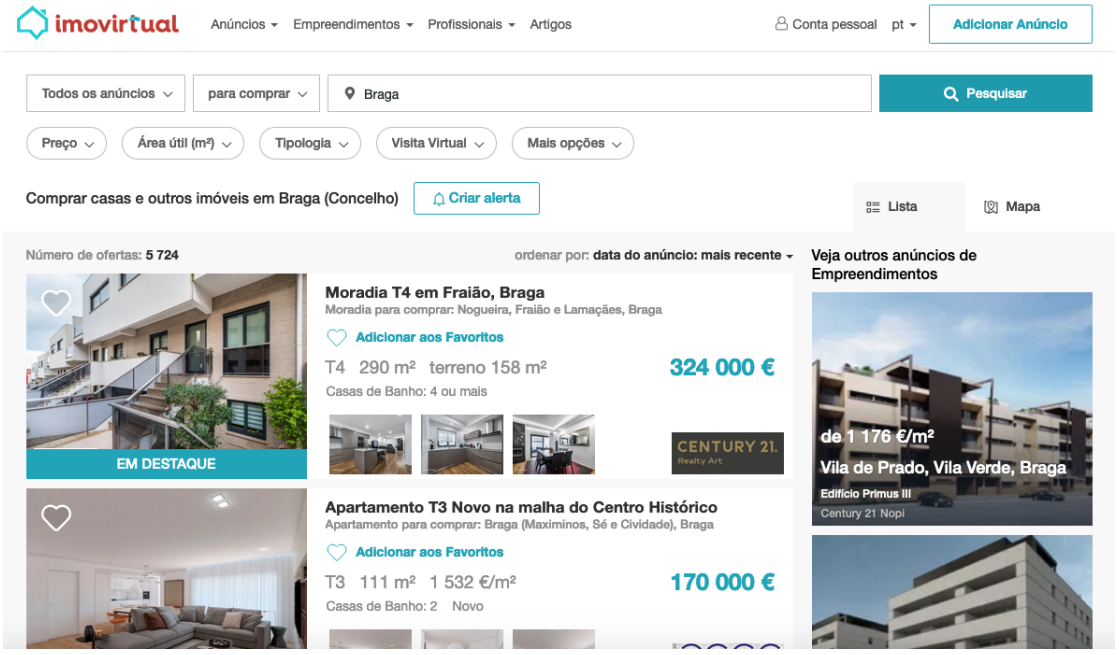


Figura 11: Visão geral dos imóveis do site da Imovirtual

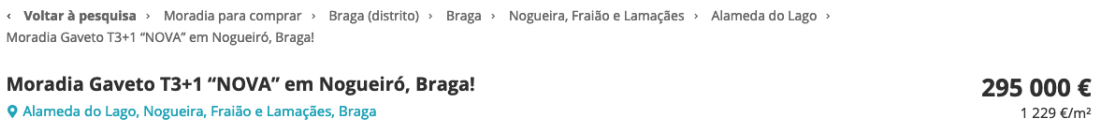


Figura 12: Características iniciais do site da Imovirtual

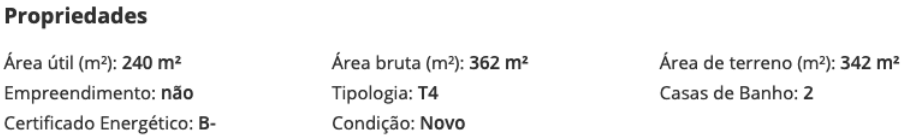


Figura 13: Características principais do site da Imovirtual

Figura 14: Características adicionais do site da Imovirtual

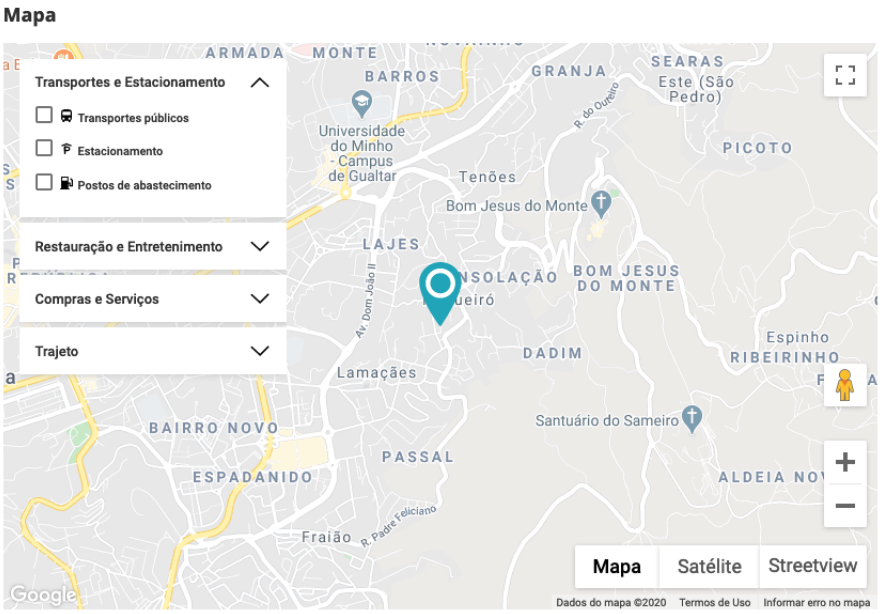


Figura 15: Coordenadas geográficas do site da Imovirtual

6.2 Algoritmo de webscraping

Para a realização da recolha de dados o nosso algoritmo tem uma sequência de passos com vista a recolher apenas os dados necessários e aqueles com os quais queremos trabalhar. Antes da construção do script é necessário inspecionar a página para identificar as classes e as seccções com características que queremos atingir com o webscraping para retirar os dados. Depois desta identificação e de perceber a página partimos para a construção do *script*.

1. Primeiro, é realizado um ciclo para realizar a recolha (descrita nos passos seguintes) para cada página de anúncios;
2. É passado o *url* do site em questão e é feito o parsing da página em questão.
3. Depois de feito o parsing, é encontrado o link para cada imóvel presente na página através do respetivo ID associado, utilizando o *soup.find_all* e passando o respetivo identificador de imóvel.
4. De seguida, após o acesso ao imóvel é feita a recolha de todas as características utilizando a função *imovelFeatures* que é descrita a seguir.
5. Como todas as características principais possuem uma determinado tag, é feito um *find* para essa tag e todas as características são colodas num dicionário *features*.
6. No caso de outras características adicionais, que possuam um ID diferente associado, é feita a procura individualmente, uma por uma.
7. No final toda esta informação recolhida no dicionário é colocada num ficheiro *.csv* respetivamente titulada.

Para a realização dos nossos scripts foram utilizadas algumas técnicas de webscraping, principalmente utilizando as funções oferecidas pelas bibliotecas *BeautifulSoup*. Como foi referido na descrição do algoritmo as duas funções mais utilizadas que nos permitem encontrar a informação que queremos são a *soup.find* e a *soup.find_all* através da passagem como pârametro de IDs ou classes associados a cada imóvel na página.

7 Pré-processamento e análise dos dados

Após recolha de informação com técnicas de *webscraping*, surge a necessidade de preparar os dados para potenciar uma boa eficiência de execução nos algoritmos de *Machine Learning*. Nesta secção é, portanto, descrito o processo realizado para preparação dos dados, assim como uma análise detalhada dos mesmos.

7.1 Leitura dos dados

Como o processo de *webscraping* aconteceu ao longo de várias semanas, foi obtido um número relativamente elevado de ficheiros *csv* contendo os dados correspondentes aos dias da semana em que foram recolhidos. Por esta razão, houve a necessidade de juntar todos esses ficheiros para formar um só *dataset* contendo toda a informação recolhida até ao momento. Assim, começamos por criar uma função que, para cada ficheiro existente, junta tudo num *dataframe* removendo os repetidos. Tal é possível pois existe um campo "Id" que permite comparar cada imóvel e assim obter um só *dataframe* com a toda a informação recolhida até ao momento sem qualquer repetição, mantendo sempre o último registo.

7.2 Processamento geral

Antes de passarmos para a análise preditiva propriamente dita, é necessário realizar um pré-processamento dos dados provenientes do *webscraping*, uma vez que muitos destes têm parâmetros vazios, informação desnecessária ou até descontextualizada. O primeiro fator a ter em atenção foi que todos os dados pertenciam a uma determinada categoria (apartamento, casa, terreno, entre outras) e, por isso, iria ser necessário dividir todo o *dataset* nas diferentes classes já referidas. Assim, depois da leitura dos dados recebidos, procedemos a um pré-processamento que seria comum a todas as instâncias que estivessem presentes. Como alguns dos parâmetros recebidos no *webscraping* eram semelhantes foi necessário juntar a informação dessas mesmas colunas bem como fazer uma normalização dos preços dos imobiliários. A adição de modificações genéricas e atribuição de tipos a valores numéricos são exemplos do processamento comum a todo o *dataset*.

A partir deste ponto, tivemos de separar os dados nas respetivas categorias e proceder então à análise e exploração dos dados. É de realçar que, mesmo tendo sido realizada a separação já referida, os procedimentos tomados a seguir foram relativamente semelhantes, com algumas diferenças nos resultados, e consequentemente, iremos seguir o *workflow* seguido nas "moradias" referenciando, sempre que necessário, diferenças significativas entre outros *scripts*.

Excluindo as colunas que possuem muita falta de informação, as *features* extraídas inicialmente dos ficheiros de dados foram as seguintes:

- Id
- Preço
- Preço m/2
- Freguesia
- Latitude
- Longitude
- Tipo de imóvel
- Tipologia
- Área útil m/2
- Condição
- Empreendimento
- Nº Casas de Banho
- Certificado energético
- Ano construção
- Área bruta m/2
- Garagem box
- Jardim
- Varanda
- Ar condicionado
- Vista de cidade
- Aquecimento central
- Video Porteiro
- Suite
- Gás canalizado
- Alarme
- Lareira

7.3 Análise exploratória dos dados

Primeiro, começamos por fazer uma análise da distribuição dos preços dos dados assim como uma visualização dos *outliers* existentes através da utilização de um diagrama de extremos e quartis e uma remoção manual de alguns pontos que achamos completamente fora do contexto e que seriam prejudiciais numa previsão de valores futura. Por exemplo, nos histogramas obtidos nas figuras 16 e 18 foi possível concluir que existe um espaçamento muito grande entre os valores abaixo e acima de 1 milhão de euros (1000000). Devido a esta verificação visual, os dados acima deste valor foram considerados *outliers* e, de forma a não enviesar a normalidade dos dados, foram retirados. A visualização dos dados após alteração é apresentada nas figuras 17 e 19. Nestas é possível verificar uma diferença menor entre os preços, estando assim mais próximos da normalidade e não de excecões. Retirando estes casos excecionais, os dados vão proporcionar melhores desempenhos na aplicação de técnicas de Machine Learning numa fase mais avançada.

Dos gráficos observa-se que, após remoção dos *outliers* iniciais, a maior gama de moradias se localiza entre os 200 mil e os 350 mil euros.

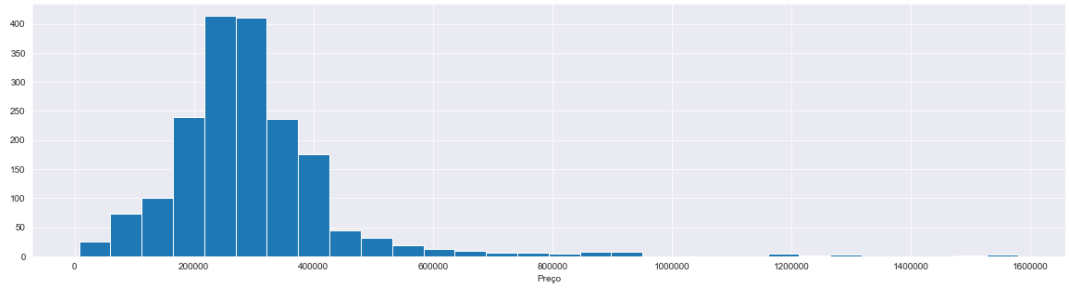


Figura 16: Histograma com o número de ocorrências dos preços, antes da remoção de outliers

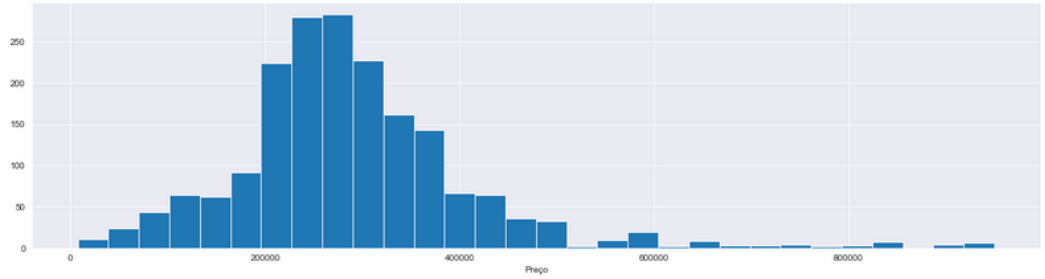


Figura 17: Histograma com o número de ocorrências dos preços, após remoção de outliers

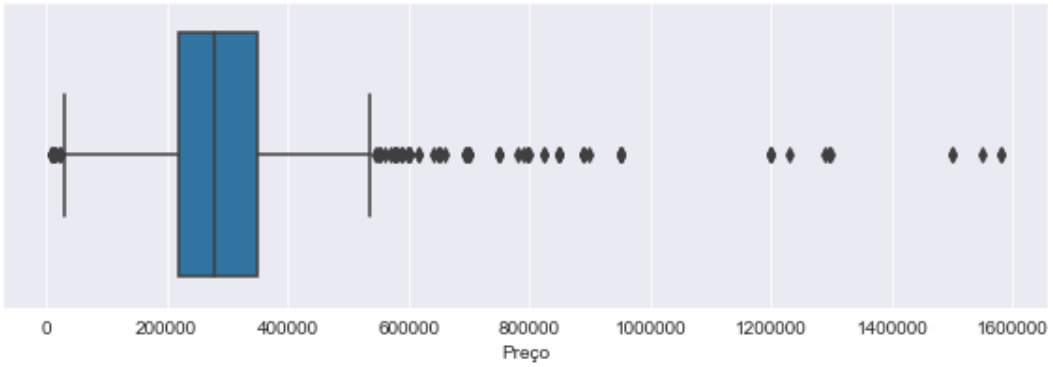


Figura 18: Diagrama de extremos e quartis do preço, antes da remoção de outliers

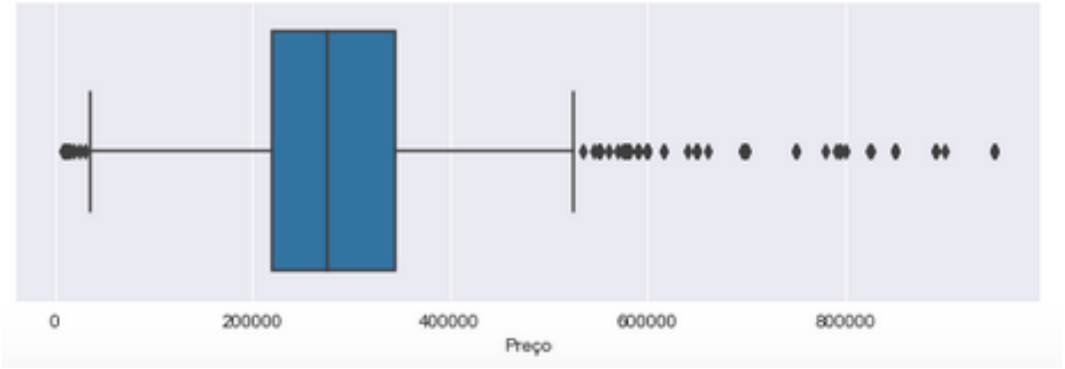


Figura 19: Diagrama de extremos e quartis do preço, após remoção de outliers

De seguida foram realizadas análises para as *features* que se apresentavam como tendo maior influência no preço: área, localização e tipologia. Devido a tal, inicialmente foi realizada a relação entre o preço e a área útil de cada imóvel (para outras categorias foi usada a área total), sendo que são mostradas as alterações realizadas nos dados nas figuras 20 (antes) e 21 (depois). Neste caso específico, os dados removidos foram os que apresentavam uma área útil superior a 10000 m^2.

Na segunda imagem é possível visualizar mais alguns casos que visualmente se verificam fora da normalidade. No entanto, nesta fase, o grupo decidiu retirar apenas os que eram evidentes, e adiar para uma fase mais avançada a análise desses casos, com recurso a técnicas próprias de análise de *outliers*.

Do gráfico pós-remoção dos *outliers* iniciais, não contabilizando as moradias com mais de 1000 m^2, verifica-se que existe uma leve tendência crescente do preço com o aumento da área.

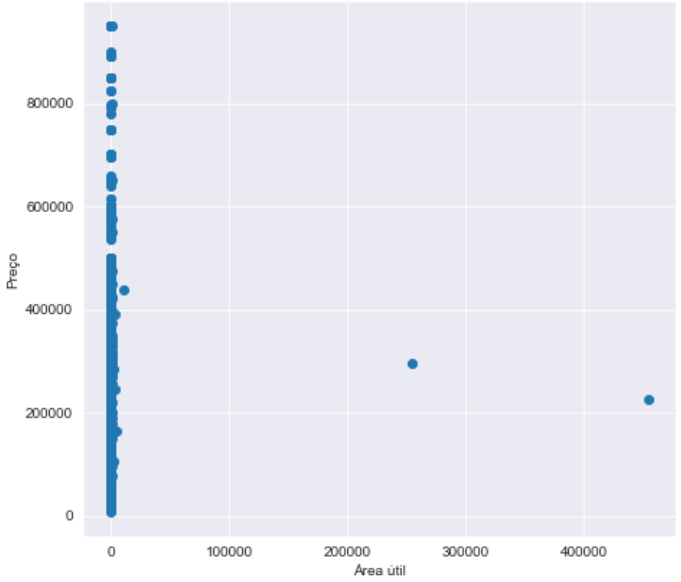


Figura 20: Relação entre o preço e a área útil de cada imóvel, antes da remoção de outliers

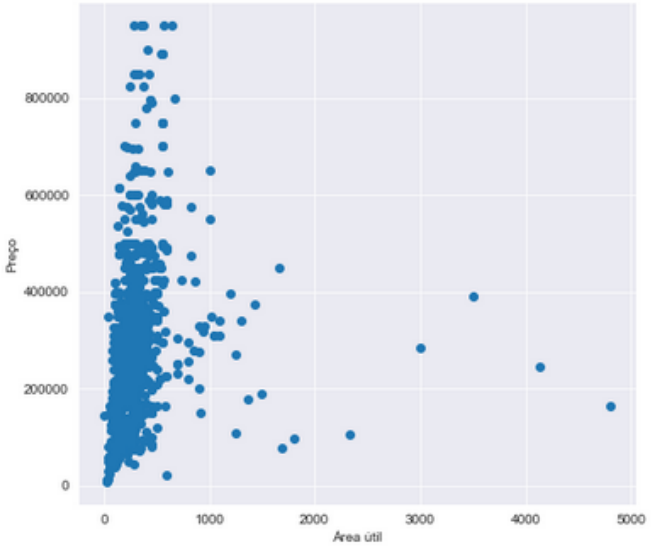


Figura 21: Relação entre o preço e a área útil de cada imóvel, após remoção de outliers

Foi também realizada uma análise do preço dos imóveis mediante a freguesia onde se encontram, através de vários diagramas de extremos e quartis. Para o caso específico das moradias, foi realizada também uma análise do preço dos imóveis mediante a tipologia do mesmo (quer seja T0, T1, T2, etc). Para estas duas *features* não foram aplicadas alterações nos dados porque a presença de *outliers* não se apresentou como tão evidente. As figuras 22 e 23 mostram a distribuição dos dados no caso particular destas *features*. Estas foram analisadas porque nesta fase se apresentavam como muito importantes, visto se relacionarem com a localização e o número de quartos dos imóveis, duas características relevantes nos preços finais.

Em relação às freguesias, visualiza-se que as moradias com valor mais elevado estão presentes nas freguesias de São Victor e Gualtar. Pelo lado contrário, os valores menores estão presentes na união de freguesias de Real, Dume e Semelhe. Nas tipologias, observa-se um aumento gradual do preço médio entre as moradias com um quarto é as moradias com seis quartos, o que é compreensível devido a todas as condicionantes que a presença de mais quartos trazem a um imóvel. Uma razão para explicar a o facto de que os T0 e T7+ não seguem de forma tão linear esta lógica é a presença de muitos poucos casos com este número de quartos. Isto indicia que estas tipologias serão muito provavelmente consideradas *outliers* na fase própria de verificação.

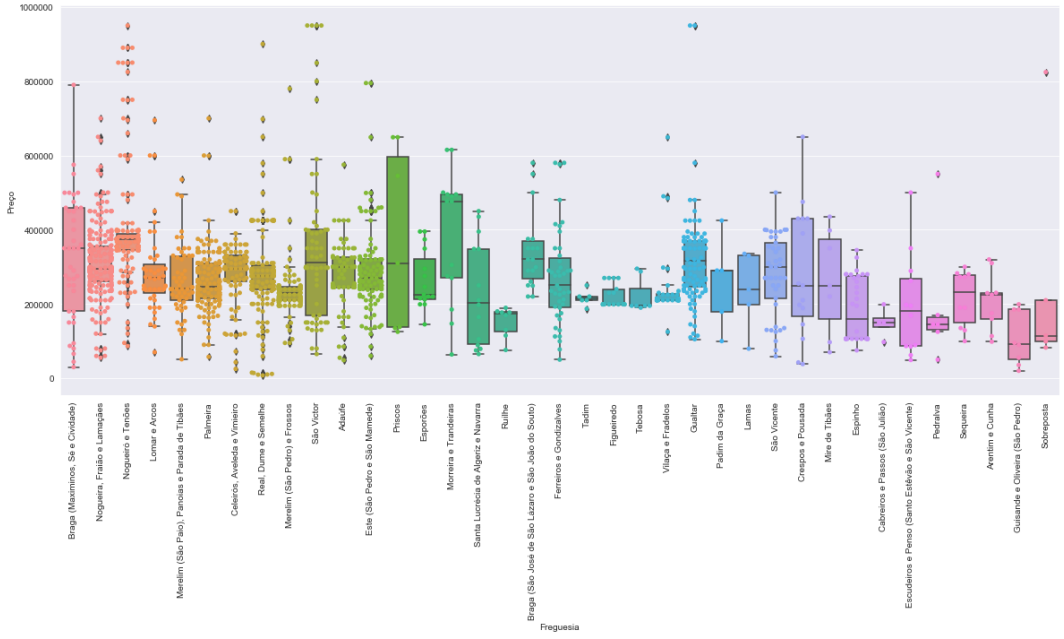


Figura 22: Diagramas de extremos e quartis das freguesias existentes

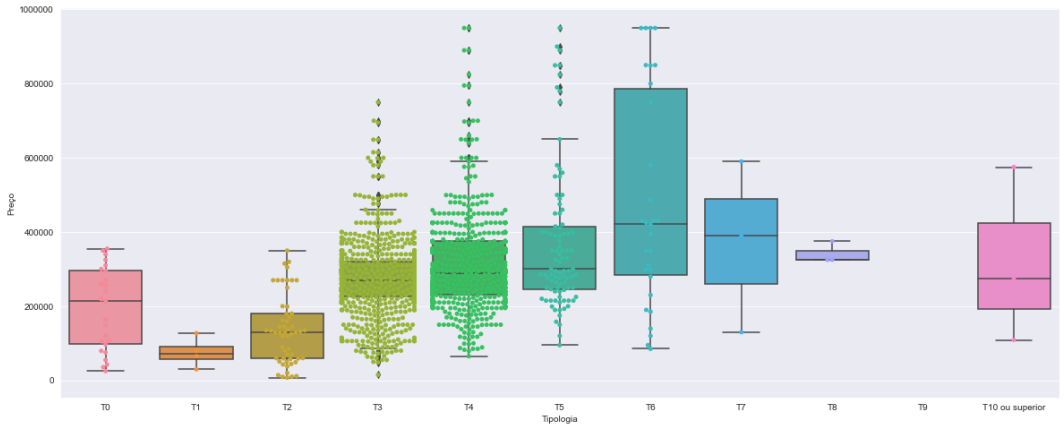


Figura 23: Diagramas de extremos e quartis das tipologias existentes

7.4 Tratamento de dados em falta

Também reparámos que existiam muitas colunas que continham uma alta percentagem de valores em falta e era necessário remove-los, uma vez que, havendo a necessidade de serem removidos ou preenchidos, sendo que, indo pela segunda hipótese inclui um risco muito elevado de inserção de informação falsa e desajustada na maior parte das entradas do *dataset*, a primeira opção é mais adequada apesar de não ser ideal. O *threshold* inicial de remoção de colunas foi de 20%, ou seja, qualquer coluna com 20% ou mais de valores em falta é removida do *dataframe*. O mesmo processo foi feito para as entradas do *dataset* e o *threshold* definido foi de 50%. Qualquer valor em falta é preenchido com a mediana dessa mesma coluna, caso seja um valor numérico, ou pela moda, para outros tipos de objetos.

	percent_missing (%)		
Garagem box	55.737705	**Tipologia**	0.000000
Área bruta m/2	53.728186	**Tipo de imóvel**	0.000000
Ano construção	52.934955	**Longitude**	0.000000
Certificado energético	52.617663	**Latitude**	0.000000
Nº Casas de Banho	38.339503	**Freguesia**	0.000000
Empreendimento	35.430989	**Preço m/2**	0.000000
Condição	21.470122	**Preço**	0.000000
Área útil m/2	0.000000	**Id**	0.000000

Figura 24: Percentagem de dados em falta nas principais colunas

7.5 Inserção de informação geográfica

Um dos fatores avaliado como essencial e que não era fornecido através do *webscraping* era a envolvência do imóvel, ou seja, os pontos estratégicos e importantes que existissem nas redondezas de cada imóvel. Sejam hospitais, centros comerciais ou escolas, foi realizada uma procura, no concelho de Braga, pelas coordenadas de infraestruturas que, a nosso ver, pudessem aumentar

ou diminuir o valor de um determinado imóvel. Após recolha dos dados foram calculadas todas as distâncias em relação a cada imóvel. Os pontos extraídos e acrescentados aos dados foram os seguintes:

- Centro da cidade
- Hospitais
- Centros de Saúde
- Centros Comerciais
- Escolas
- Universidades
- Estação de Comboios
- Parques Industriais
- Central de Autocarros
- Bancos
- Correios
- Parques e zonas verdes
- Serviços Públicos
- Polícia Municipal
- Farmácias

7.6 *Outliers* e *encoding*

Devido à adição das distâncias geográficas aos pontos mais importantes do concelho, foi obtida uma coleção de dados dotada com as informações tidas como essenciais para a avaliação de um imóvel. Todas as *features* utilizadas nos modelos de regressão estavam já recolhidas. Entre as *features*, existiam algumas que eram apenas identificadores e não características do imóvel, neste caso: `Tipo de imóvel`, `Longitude` e `Latitude`. Foram, por isso, retirados da coleção de dados. Foi também retirada a *feature* `Preço m/2` pois o `Preço` já estava a ser tido em conta na coleção de dados.

Nesta fase, foi realizada uma remoção de *outliers* de forma mais pormenorizada, ao contrário da inicial que foi mais básica. Como tinha sido visualizado nas figuras 17 e 19, existe uma série de *outliers* com valores de preço superior a 500 mil euros e outros com valores inferiores a 50 mil euros. Esses foram, portanto, retirados, dando resultado na seguinte distribuição apresentada na imagem 25. Nesta imagem é possível observar uma maior ocorrência de imóveis com preço entre os 200 mil e os 350 mil euros, como já se tinha concluído, mas agora todos os restantes valores possuem casos suficientes para não serem considerados *outliers* e assim não prejudicarem a análise.

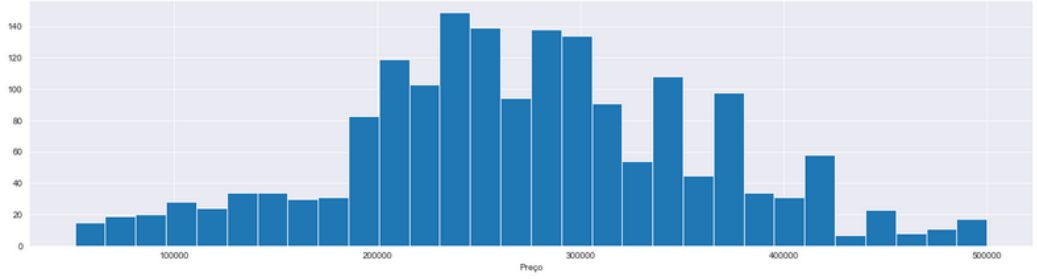


Figura 25: Histograma com o número de ocorrências dos preços

Para a área útil de cada imóvel foi realizado o mesmo processo, tendo sido retirados os imóveis com áreas superiores a 800 m^2, pois pela figura 21 era conclusivo que estes valores estavam

longe do *cluster* principal observado. Esses foram, portanto, retirados, dando resultado na seguinte distribuição apresentada na imagem 26.

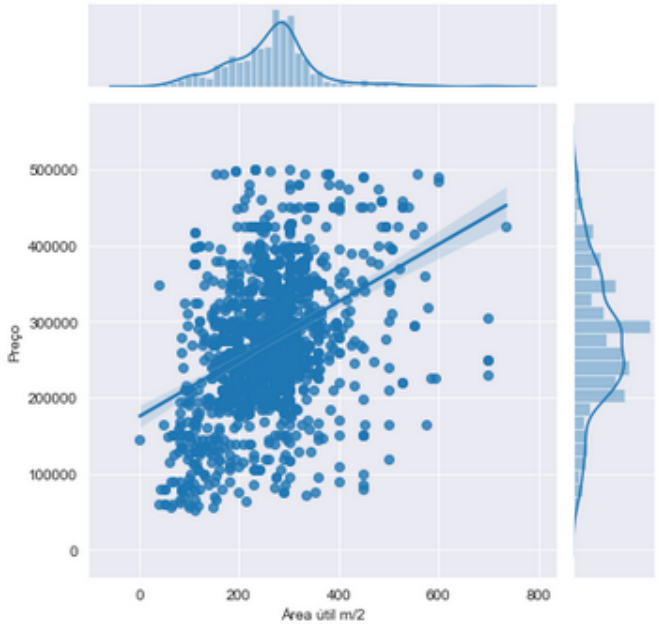


Figura 26: Relação entre o preço e a área útil de cada imóvel

Em relação às freguesias, apesar de algumas apresentarem poucos dados, visualmente não foi realizada nenhuma remoção de *outliers*. Na *feature* tipologia é possível visualizar um aumento de preço consoante o aumento do número de quartos, o que faz sentido pois o número de quartos é, idealmente, o principal fator diferenciador nos preços das casas. Devido a tal, e tendo em conta que quantos mais quartos mais o preço seria elevado, foram retiradas as tipologias que não seguiam esta lógica, sendo assim considerados como *outliers*. Foram também consideradas como *outliers* as tipologias que possuíam muitos poucos casos. Nesta fase já tinham sido identificados e retirados outros *outliers* e, portanto, a distribuição era a apresentada na imagem 27. A menor frequência de casos nos extremos mostra a coincidência da consideração de *outliers* noutras *features*, que também se sentiram nesta. A distribuição, após remoção, era a apresentada na imagem 28.

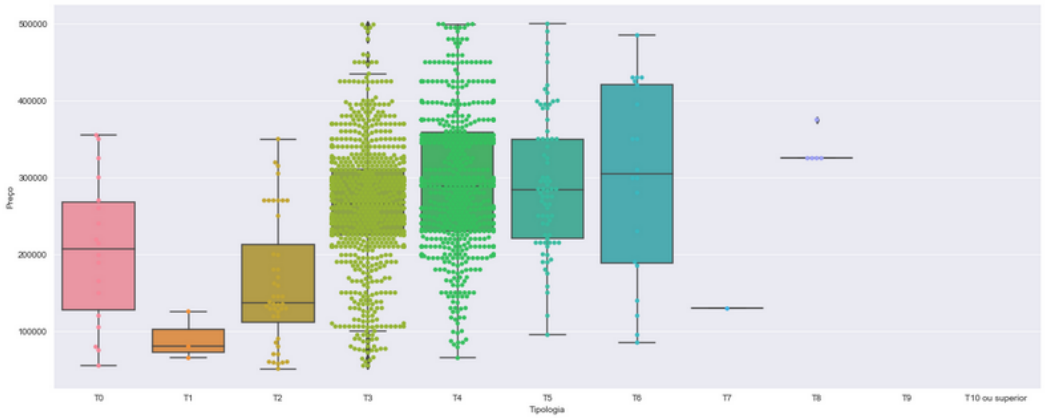


Figura 27: Diagramas de extremos e quartis das tipologias existentes, antes da remoção

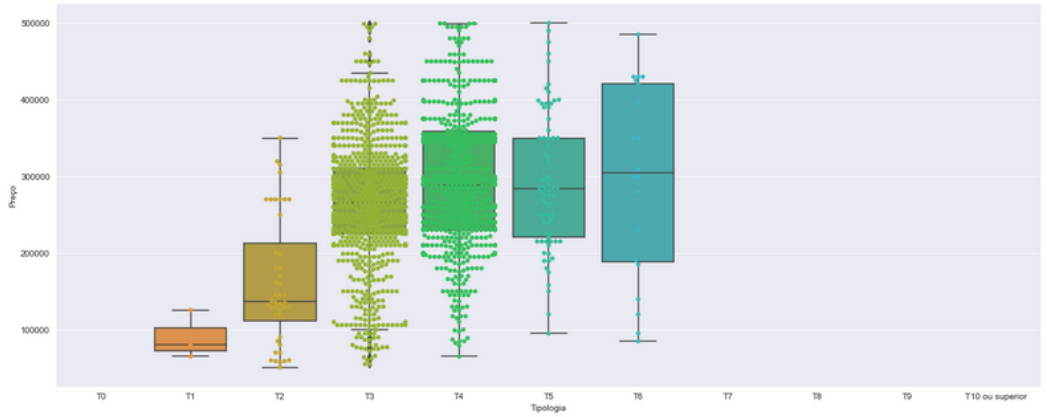


Figura 28: Diagramas de extremos e quartis das tipologias existentes, após remoção

Como existia a noção que iriam ser utilizados vários tipos de modelos para prever os preços dos imóveis, foi tido como importante nesta altura fazer o *encoding* de atributos categóricos quer usando *label encoding* quer usando *one hot encoding*, uma vez que diferentes modelos exigem tratamentos diferentes deste tipo de atributos. Desta forma, atributos categóricos passam a valores numéricos, pois a generalidade dos modelos não funcionam com valores categóricos.

Após *encoding* dos dados, foi possível realizar remoção dos dados numéricos de forma automática. Decidimos usar um método matemático que calcula o valor *zscore* que, dado um conjunto de valores, subtrai um dos valores com a sua média e divide essa operação pelo desvio padrão e, assim, é possível obter os *zscore* de todos os valores do *array* correspondente. Assim, através de um valor de *threshold*, retiramos todas as linhas do *dataframe* que possuam *zscores* mais elevados que o limite por nós definido. Isto permite-nos ter também uma distribuição de preços que se aproxima mais de uma curva normal o que, à partida, terá um impacto positivo na previsão dos mesmos, uma vez que este processo é equivalente a retirar valores que, numa distribuição normal, estariam nos extremos, o que levará a uma estandardização dos nossos dados que, à partida, terá um impacto positivo nos modelos de regressão.

$$z = \frac{x - \mu}{\sigma}$$

Figura 29: Fórmula para cálculo do *zscore*

Além duma pormenorização nos *outliers* já identificados visualmente, este método realizou também uma identificação e remoção de *outliers* nas restantes *features* dos dados, que não seriam tão interpretáveis visualmente. Desta forma, o método matemático proporciona uma boa solução.

7.7 Simetria dos dados

Por fim, foi realizada uma análise de simetria dos dados que nos mostrou que estes estavam bastante bem distribuídos em relação a uma função normal, como podemos ver nas figuras seguintes. Foram aplicadas técnicas de ajuste dos dados, mas estes não se tornaram mais ajustados à normal.

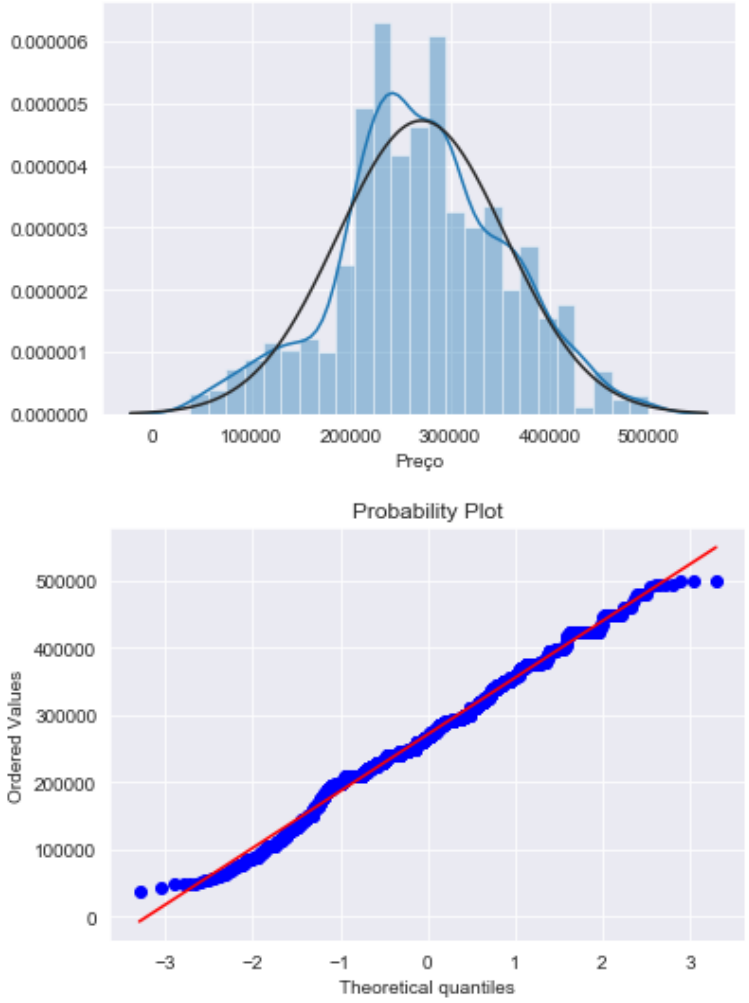


Figura 30: Distribuição dos preços em comparação com uma normal

7.8 Análise de correlação

Depois de analisados e tratados todos os dados, procedeu-se a uma análise de correlação entre as colunas do *dataset*. Isto permite identificar aspetos do mesmo que transmitam aproximadamente a mesma informação. Na seguinte matriz, é possível verificar os valores de correlação entre todas as colunas do *dataframe*, valores estes que estão entre -1 e 1, sendo que valores mais pertos de 1 significam relações positivas (o aumento do valor de uma variável leva ao aumento da outra), valores mais pertos de -1 significam relações negativas (o aumento do valor de uma variável leva à diminuição da outra) e valores mais pertos de 0 significam que a variação do valor de uma variável não tem influencia no valor da outra.

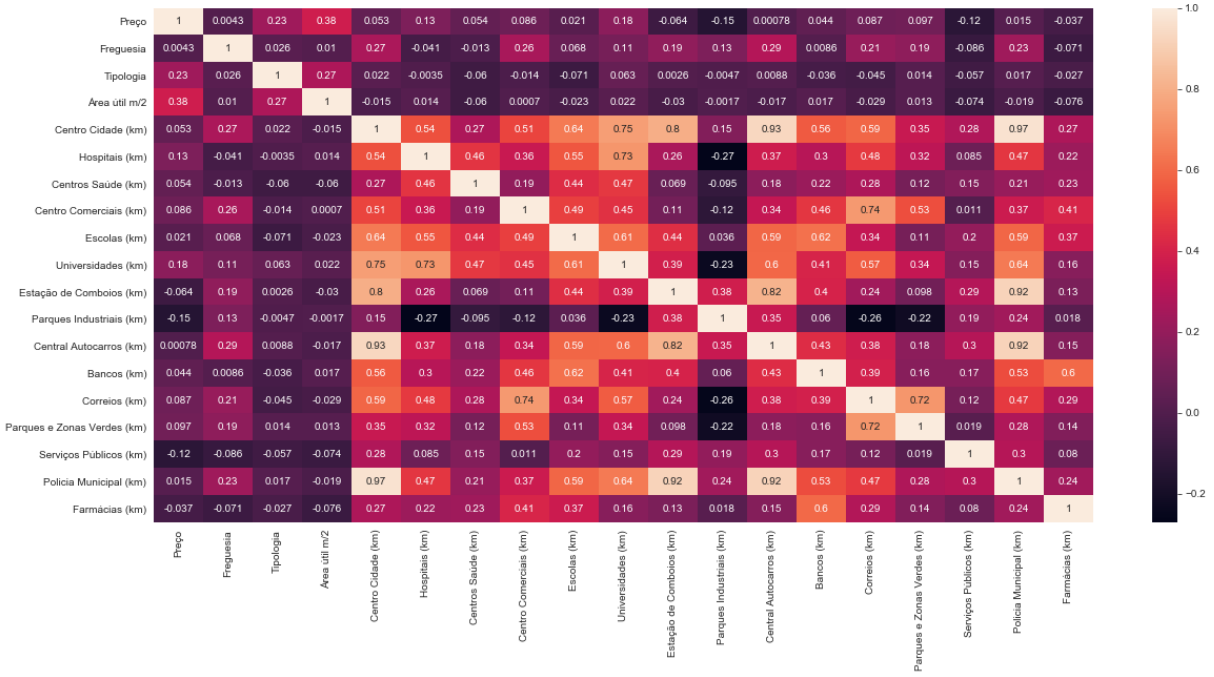


Figura 31: Matriz de correlação

Da matriz é possível concluir que as *features* com mais influência positiva no preço são:

- Área útil (m^2): 38%
- Tipologia: 23%
- Proximidade a universidades: 18%
- Proximidade a hospitais: 15%

Por outro lado, as *features* com mais influência negativa no preço são:

- Proximidade a parques industriais: -15%
- Proximidade a serviços públicos: -12%

Em relação às *features* positivas, é percetível que a área e a tipologia sejam os que possuem mais influência, pois são estes os principais diferenciadores entre imóveis. A proximidade a universidades e hospitais é também sentida positivamente pois estas infraestruturas são escassas no concelho. Por outro lado, os serviços públicos estão mais presentes em várias freguesias e, apesar de importantes, correlacionam o preço com o facto de estarem presentes em freguesias com baixas gamas de preços. Os parques industriais, de forma semelhante, estão também localizados nos subúrbios e, devido a tal, correlacionam-se com a baixa de gama de preços de imóveis do concelho.

8 Aplicação de técnicas e algoritmos de regressão

A partir deste momento, podemos afirmar que os nossos dados estão prontos para passar para a fase da modelação. Na maior parte dos modelos de *machine learning* é necessário dividir os dados nas variáveis a estudar (X) e a variável resposta (y que corresponde ao preço), bem como fazer uma divisão entre dados de treino e de teste que serão valiosos para a avaliação de cada modelo apresentado a seguir.

8.1 *Linear Regression*

Para começar, optamos por utilizar um modelo mais simples utilizando o módulo *linear_model* dentro da *API* do *scikit-learn*, o *Linear Regression*. Para a sua criação, apenas foi necessário utilizar o método sem parâmetros adicionais porque havendo já de início poucos elementos para serem alterados, não achamos necessário alterar o que já está predefinido nessa função. Depois de feitos o treino (com os dados de treino) e o teste (com os dados de teste) fizemos a avaliação do modelo tendo em conta alguma métricas, nomeadamente, o *mean absolute error* (MAE) que faz a computação da média do erro absoluto obtido entre os dados reais e os previstos, o *mean squared error* (MSE) que faz a computação da média do erro quadrático entre os dados reais e previstos e o *root mean squared error* (RMSE) que faz a computação da raiz da média do erro quadrático entre os dados reais e os previstos. Em baixo, apresentamos os gráficos obtidos dos resultados provenientes do modelo em que as métricas calculadas tomam os seguintes valores:

- MAE → 58150.01 €
- MSE → 5935972863.09 €
- RMSE → 77045.26 €

Figura 32: Gráfico da diferença entre preços reais e previstos

Figura 33: Gráfico de dispersão dos preços reais e previstos

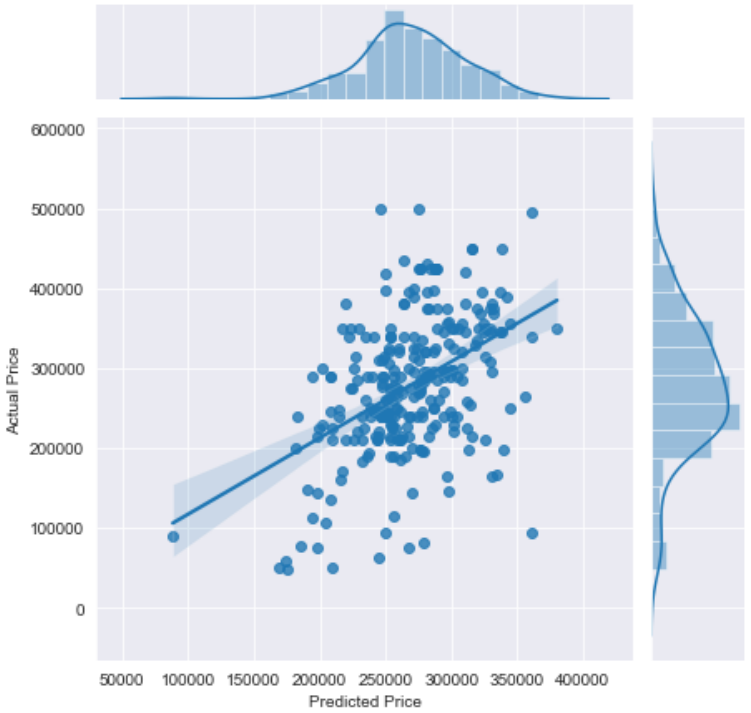


Figura 34: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.2 *Random Forest*

Também optamos por tentar um modelo de *random forest* que consiste em criar um número elevado de árvores de decisão que atuam em porções de dados escolhidos aleatoriamente, o que permite diminuir a variância do modelo. Isto deve-se ao facto de árvores diferentes não usarem os mesmos dados o que também permite uma baixa correlação entre elas, tendo posteriormente consequências positivas para o nosso modelo.

Falando agora da parte mais prática, utilizamos mais uma vez a *API* do *scikit-learn* mas com o módulo *ensemble* que contem o método "RandomForestRegressor". Este contém um número de parâmetros muito mais elevado e que podem ter grande influência nos resultados do modelo, por isso tivemos também que utilizar algo que nos permitisse achar os melhor parâmetros para essa função. Decidimos optar por usar uma estratégia de *grid search* que consiste e fornecer, em estilo dicionário, os parâmetros que pretendemos testar e acabaremos por saber qual a combinação ou combinações que resultam nos menores valores de erro. De realçar que este método existe no módulo *model_selection* da *API* do *scikit-learn*. Em baixo, apresentamos os gráficos resultantes do treino e teste deste modelo bem como as métricas calculadas.

- MAE → 46231.68 €
- MSE → 4127766959.88 €
- RMSE → 64247.70 €

Figura 35: Gráfico da diferença entre preços reais e previstos

Figura 36: Gráfico de dispersão dos preços reais e previstos

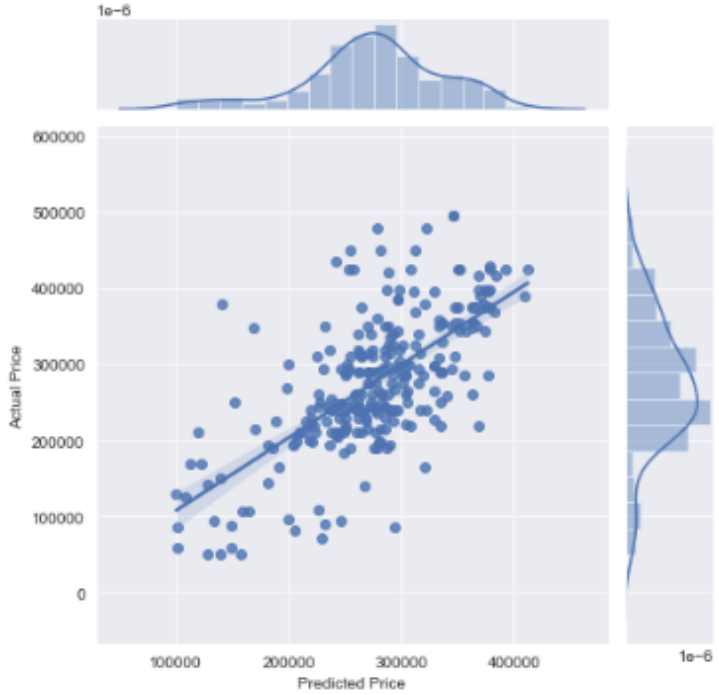


Figura 37: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.3 *Ridge Regression*

O problema da multicolinearidade é um problema comum nos modelos de regressões em que variáveis independentes possuem relações linearas iguais ou muito aproximadas umas das outras. Assim, surge este modelo de regressão, o *Ridge Regression*, especificamente útil para mitigar este problema que surge com maior frequência em problemas com elevado nº de parâmetros.

Para a implementação, utilizamos a *API* do *scikit-learn* recorrendo ao módulo *linear-model* que contém o classe *Ridge*. Tal como no modelo anterior *Random Forest* optámos pela estratégia de *grid-search* que é um método já fornecido na classe *Ridge* não sendo necessário fazer nenhum import adicional.

- MAE → 58668.45 €
- MSE → 5643428609.07 €
- RMSE → 75122.76 €

Figura 38: Gráfico da diferença entre preços reais e previstos

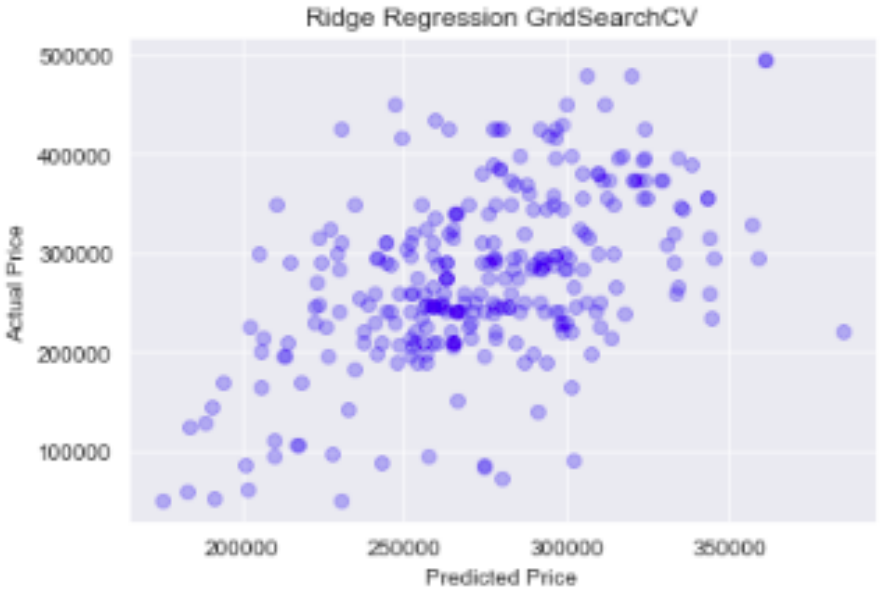


Figura 39: Gráfico de dispersão dos preços reais e previstos

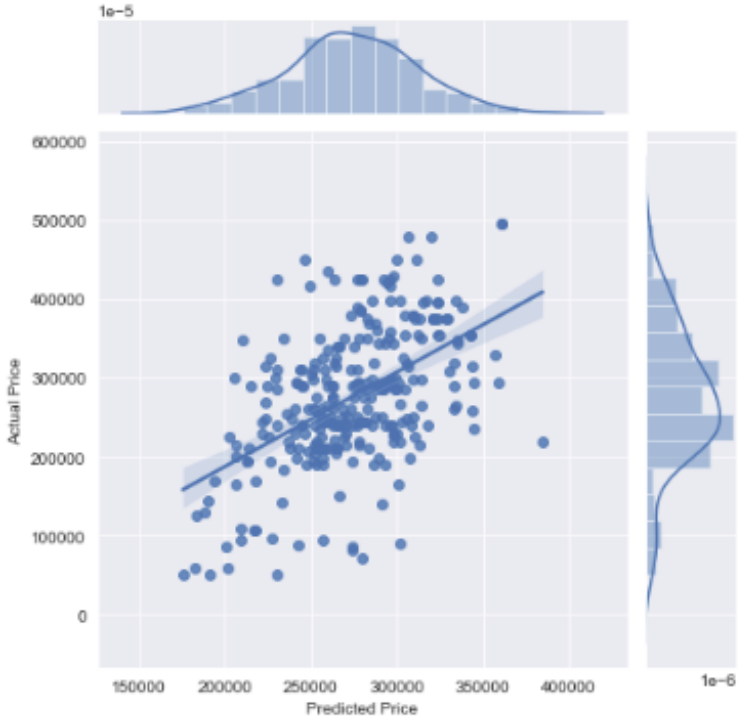


Figura 40: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.4 *Gradient Boosting Regressor*

Este modelo tem alguma parecenças com o do *random forest*. É uma técnica que produz um modelo de previsão através da junção de outros modelos mais simples, sendo as árvores de decisão bastante populares, mas que, neste caso, é criado através de iterações onde, na primeira é gerado um modelo simples e nas instâncias a seguir são treinadas usando os resíduos da iteração anterior. Por outras palavras, podemos dizer que à medida que vamos avançando nas iterações, iremos ter modelos que conseguem ter valores de erros cada vez menores, uma vez que, como já referimos, cada um recebe os resíduos do anterior e, por isso, aprende com os erros do modelo passado. Mais uma vez utilizamos o método *grid search* do *scikit-learn* pois havia a necessidade de otimizar os seus parâmetros. Em seguida, apresentamos os gráficos correspondentes a este modelo bem como as métricas de erro calculadas.

- MAE → 49523.69 €
- MSE → 4574469913.76 €
- RMSE → 67634.83 €

Figura 41: Gráfico da diferença entre preços reais e previstos

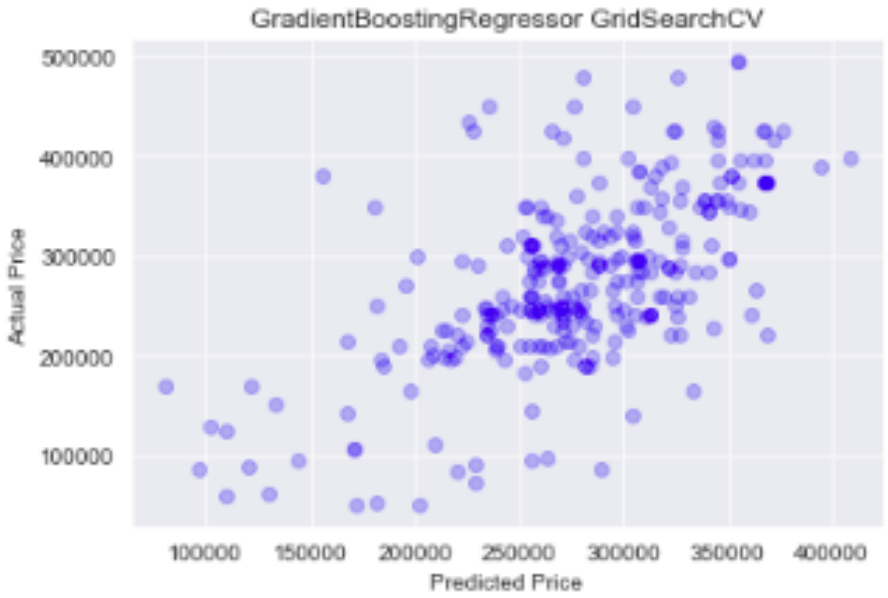


Figura 42: Gráfico de dispersão dos preços reais e previstos

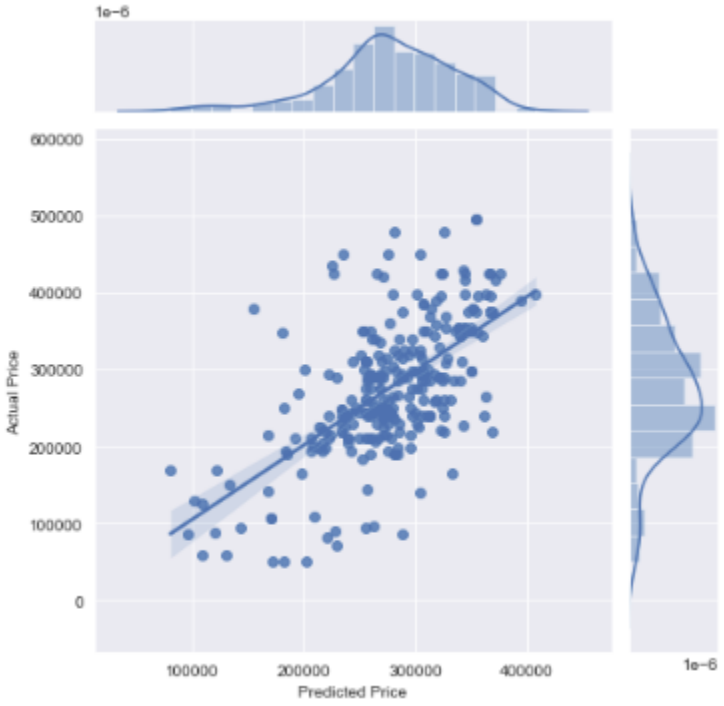


Figura 43: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.5 *Decision Tree Regression*

Uma *decision tree* constrói modelos de regressão ou classificação na forma de uma estrutura em árvore. Ele divide um conjunto de dados em subconjuntos cada vez menores, ao mesmo tempo em que uma árvore de decisão associada é desenvolvida de forma incremental. O resultado final é uma árvore com nós de decisão e nós de folha. Um nó de decisão possui duas ou mais ramificações, cada uma representando valores para o atributo testado. O nó folha representa uma decisão sobre o destino numérico. O *AdaBoost* é um algoritmo meta-heurístico, e pode ser utilizado para aumentar a performance de outros algoritmos de aprendizagem. Este algoritmo é adaptável no sentido de que as classificações subsequentes feitas são ajustadas a favor das instâncias classificadas negativamente por classificações anteriores. Em baixo, apresentamos os gráficos resultantes do treino e teste deste modelo bem como as métricas calculadas.

- MAE → 54708.03 €
- MSE → 4385318090.94 €
- RMSE → 66221.73 €

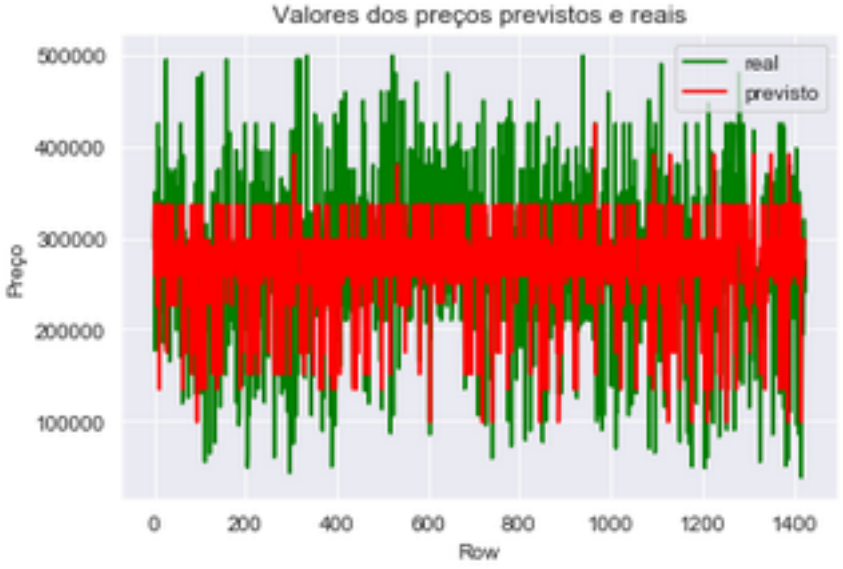


Figura 44: Gráfico da diferença entre preços reais e previstos

Figura 45: Gráfico de dispersão dos preços reais e previstos

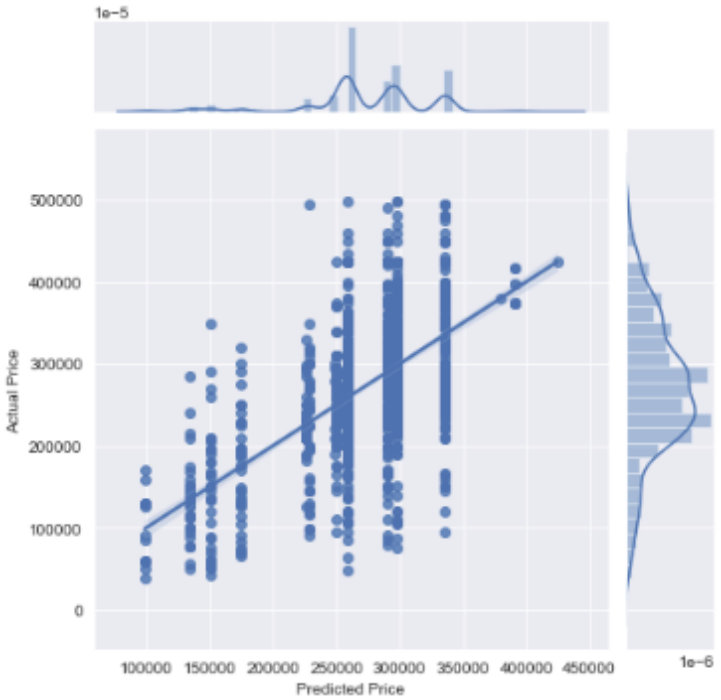


Figura 46: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.6 *LightGBM Model*

Para continuar a nossa procura por soluções diferentes e melhores decidimos optar pela mudança de API, neste caso aqui, para *LightGBM*. Este tipo de modelo é baseado em algoritmos de aprendizagem com árvores utilizando uma framework de *gradient boosting*. Este cresce a árvore verticalmente, enquanto outros algoritmos com estas bases crescem a árvores horizontalmente, o que significa que o *LightGBM* cresce a árvore em folha, enquanto outros algoritmos crescem em nível. Um dos pontos positivos deste modelo que nos levou a experimentá-lo deve-se ao facto deste suportar uma grande quantidade de dados em larga escala com uma boa precisão. O *GBM* é prefixado como "Leve"devido à sua alta velocidade. O *LightGBM* pode lidar com o grande tamanho de dados e leva menos memória para ser executado. Outro motivo pelo qual o modelo é popular é porque ele se concentra na precisão dos resultados.

Para a implementação do modelo foi necessário preparar os dados para inserir no modelo e não foram utilizados quaisquer outros métodos que não os fornecidos pela API do *LightGBM*. Os gráficos correspondentes e as métricas calculadas através deste modelo são apresentados a seguir:

- MAE → 47764.97 €
- MSE → 4398618140.95 €
- RMSE → 66322.08 €

Figura 47: Gráfico da diferença entre preços reais e previstos

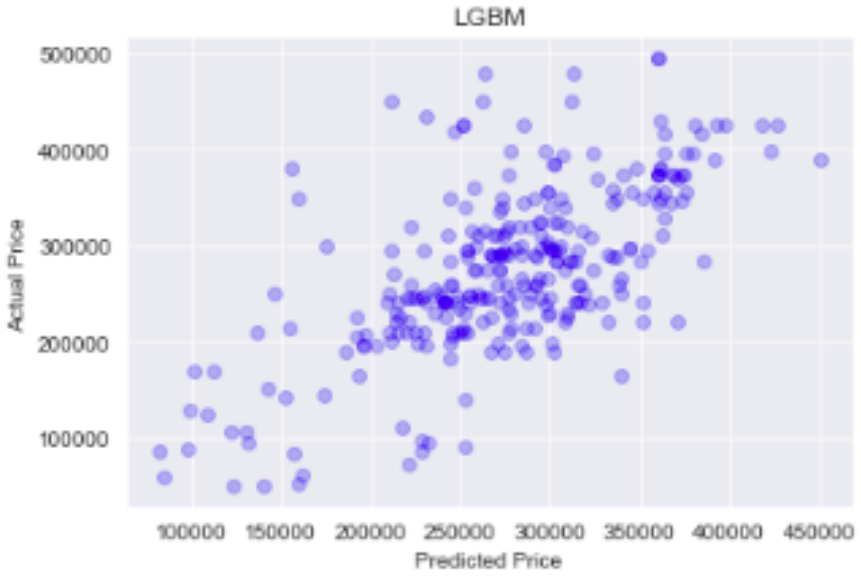


Figura 48: Gráfico de dispersão dos preços reais e previstos

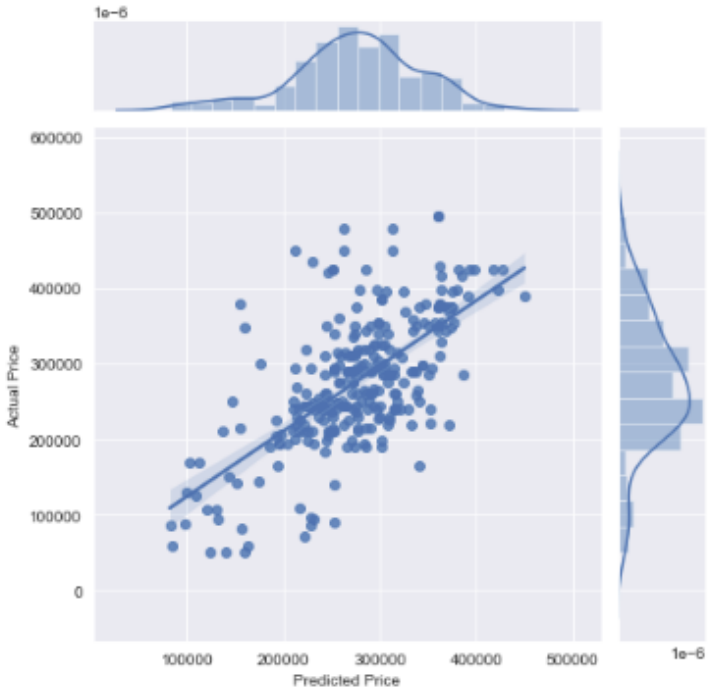


Figura 49: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

8.7 *Artifical Neural Network*

Aqui decidimos optar por usar uma *API* diferente das que já tinhamos usado para os outros modelos, chamada *keras*. Apesar de ser semelhante a uma *deep neural network* decidimos testar na mesma pois, uma vez que, neste caso, podemos também verificar, através de curvas de aprendizagem, como se processa o treino da rede e também gerir a estrutura da rede quer seja o número de camadas e de neurónios em cada camada, funções de ativação, percentagem de *dropout* entre vários outros parâmetros. Para podermos gerar as curvas de treino, é necessário que haja um conjunto de dados de validação que não pode estar nem dentro dos dados de treino nem dos de teste e, para isso, é possível usar um dos parâmetros da função "fit" que permite fazer uma partição dos dados de treino segundo um valor decimal. Além dos resultados dos gráficos apresentados em todos os modelos para análise e comparação de resultados, apresentamos também as curvas de aprendizagem da rede.

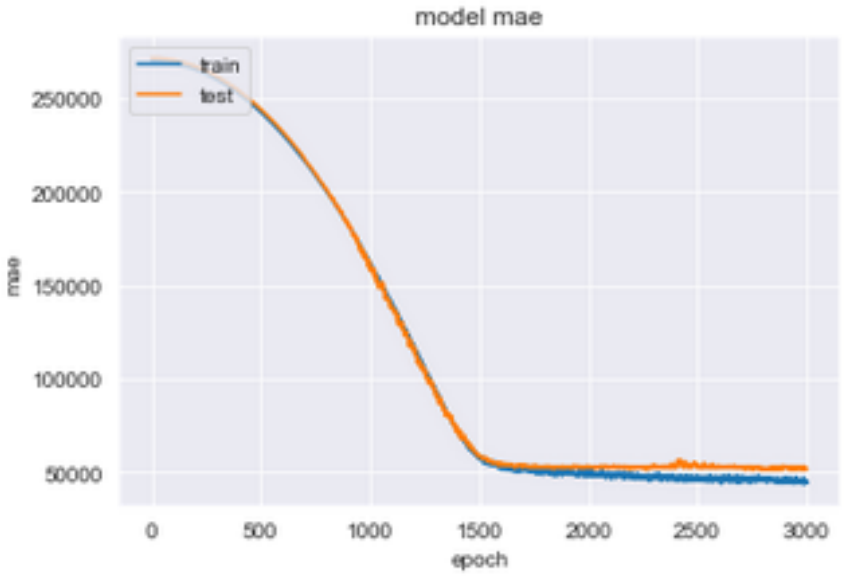


Figura 50: Loss/MAE associada ao treino da rede

Figura 51: Gráfico da diferença entre preços reais e previstos

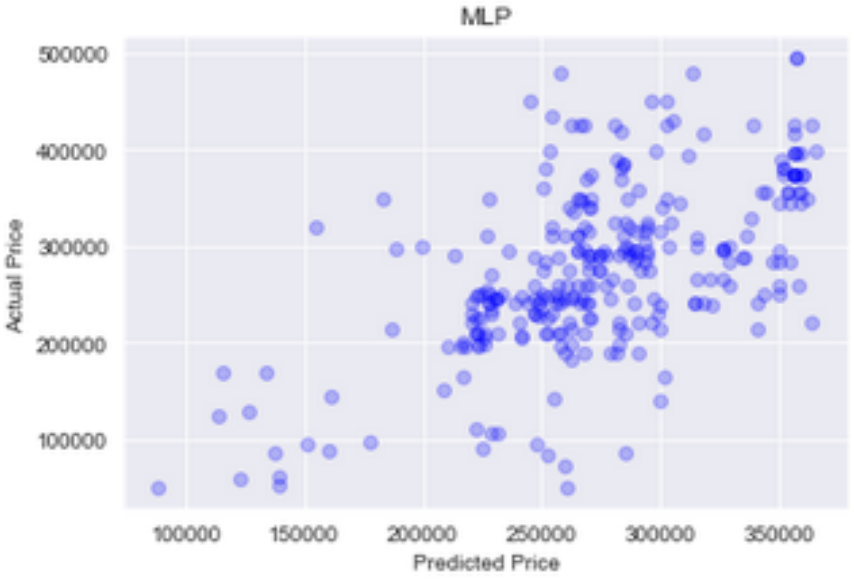


Figura 52: Gráfico de dispersão dos preços reais e previstos

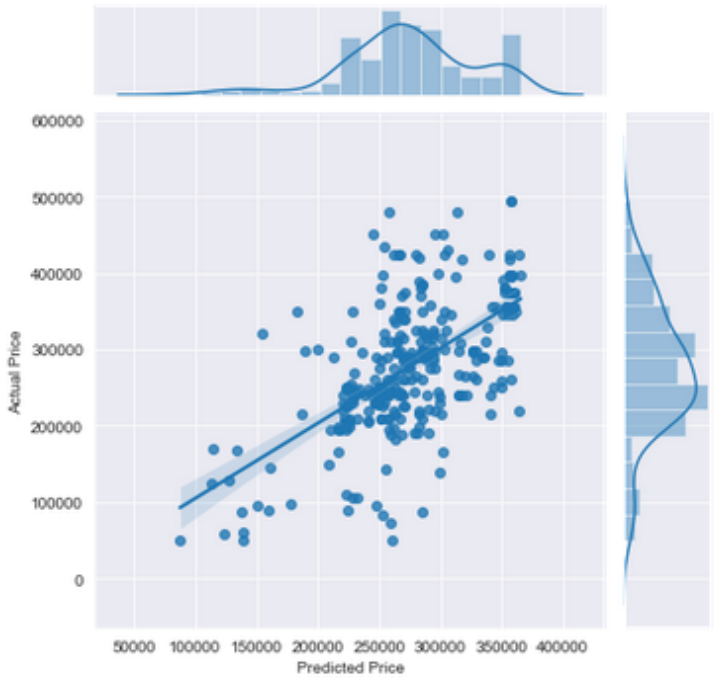


Figura 53: Gráfico de dispersão dos preços reais e previstos e sua distribuição com reta linear

9 Análise da regressão

Após processamento dos dados e aplicação dos modelos apresentados na secção anterior, surge a necessidade de analisar os resultados obtidos e de escolher o modelo que proporciona melhor precisão para a regressão. Para essa comparação vai ser utilizada a métrica MAE (erro médio absoluto). As outras métricas, nesta fase, não são tidas em conta porque os seus significados não são tão perceptíveis. Com a utilização do MAE a comparação entre modelos torna-se compreensível. A tabela seguinte mostra os valores referenciados para cada um dos modelos testados.

Modelo	**MAE**
Linear Regression	58150.01 €
Random Forest Regression com GridSearchCV	46231.68 €
Ridge Regression com GridSearchCV	58668.40 €
GradientBoostingRegressor com GridSearchCV	49523.60 €
Decision Tree Regression com AdaBoost	54708.03 €
LightGBM	47764.97 €
Artificial Neural Networks (3 camadas)	51729.27 €

Após observação da tabela, facilmente são observáveis três modelos que apresentaram maior precisão, isto é, os que possuem um erro médio absoluto inferior a 50 mil euros. Este erro pode ser explicado pela quantidade e variabilidade de dados que não foi possível ser muito elevada, devido às políticas de privacidade dos *websites* das imobiliárias. Um outro fator que pode estar a influenciar o erro é a desconsideração que pode estar a existir de alguns fatores por falta de informação ou até mesmo irregularidades no pré-processamento dos dados. Além destas razões, também é influenciador o facto do mercado imobiliário possuir bastantes e variados interesses entre cada vendedor, o que faz com que exista essa difícil tarefa de quantificar o interesse de cada terceiro.

Nos gráficos apresentados na secção anterior, para cada um dos três melhores modelos, é possível observar a capacidades destes para reconhecer casos mais drásticos. É principalmente com esta capacidade que estes três modelos conseguem ter um erro menor em relação aos restantes. De entre estes três modelos, qualquer um seria bem escolhido pois a diferença do erro entre estes não é substancial, mas surge também a necessidade de ter em conta o tempo de processamento e a capacidade para a quantidade de dados introduzidos. Importante fazer referência ao *GridSearch Cross Validation*, um algoritmo otimizador com base na procura dos parâmetros para melhor desempenho dos modelos. Este foi aplicado a alguns dos modelos e, devido à sua utilização, esses modelos apresentam também maior eficiência. Aliada a esta maior eficiência está um maior tempo de execução, sendo que este é um ponto negativo desses dois modelos que apresentaram um erro médio absoluto inferior aos 50 mil euros. Ao contrário desse maior tempo de execução, o *LightGBM* apresentou, igualmente, um resultado bom e uma execução substancialmente mais rápida.

Caso o único fator de decisão se concentrar no erro médio absoluto, então a melhor solução seria o Random Forest Regression com GridSearchCV. No entanto, após toda a análise e considerando vários fatores como o tempo de execução, o *LightGBM* foi o escolhido como o melhor modelo para o problema em questão.

10 Aplicação de técnicas de previsão do *target*

Depois de concluída a fase da aplicação de técnicas e algoritmos de regressão e análise dos respetivos resultados, procedemos então à fase de *forecast*. Por isso, em seguida, apresentaremos o processo desenvolvido para esse fim bem como os algoritmos usados para a previsão de valores futuros.

Naturalmente, como tínhamos os dados recolhidos diariamente, foi necessário, mais uma vez, passar pelo processo de englobar todos os dados contidos nos nossos *csvs*. No entanto, neste caso, o agrupamento dos dados foi feito de maneira diferente. Como era necessário colocar todos os preços em série e por cada identificador recolhido, decidimos criar uma estrutura dicionário em *python* que nos permitiu colocar cada identificador como chave bem como associar como valores *arrays* de datas, preços e previsões, este último a ser preenchido posteriormente. Decidimos fazer um pequeno gráfico pegando a lista de datas e de preços de um imóvel aleatório para confirmar que este processo tinha ficado concluído corretamente, descrito em seguida.

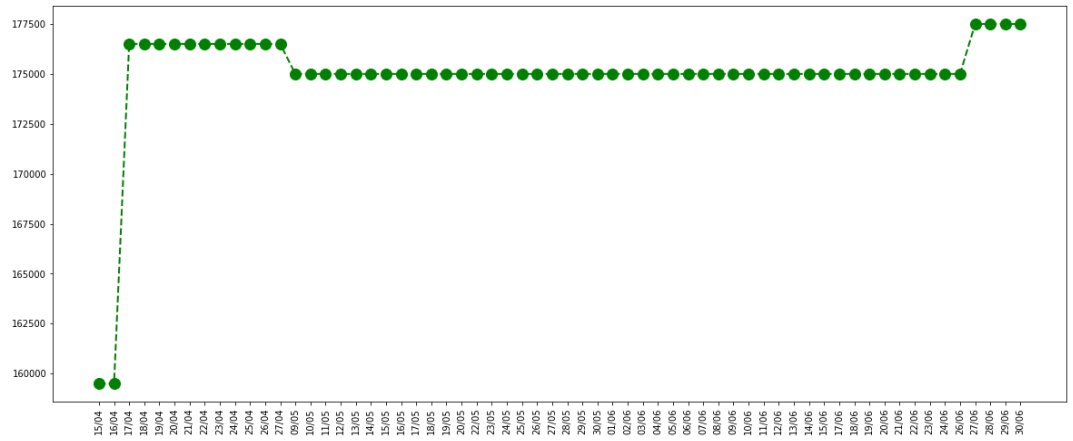


Figura 54: Gráfico do imóvel com o id 11160514 descrevendo a relação entre preço e respetiva data

Fizemos uma pequena pesquisa sobre que algoritmos de *time series forecast* usar e encontrámos principalmente 3. Primeiro começamos até por tentar usar uma rede LSTM (*Long short-term memory*) mas percebemos logo a seguir que este tipo de solução não se enquadrava com as especificidades do problema. Apresentámos então os algoritmos que implementámos para esta problemática.

10.1 *Auto Regression*

Este método permite obter o próximo passo na sequência temporal utilizando uma função linear das observações anteriores. Também é possível estabelecer um valor de ordem do modelo que permite dizer que termos anteriores entram também na previsão temporal além do cálculo normal da regressão

10.2 *Simple Exponential Smoothing*

Este método é utilizado normalmente em dados que não tenham tendência nem sazonalidade aparentes. Existem alguns métodos mais vulgares que utilizam apenas 1 ou mais (reduzidas) observações mais recentes para obter previsões de valores enquanto que outras consideram todos

os valores passados como tendo a mesma importância para fazer o *forecast*. Aqui não se aplica nenhum dos modelos anteriores, uma vez que este método considera todas as observações passadas mas atribuindo pesos (importância) diferentes a cada um deles, dando mais relevo às mais recentes, através de uma função exponencial, ou seja, à medida que percorremos os valores em direção ao passado, os seus pesos vão diminuindo exponencialmente.

10.3 *Holt Winter's Exponential Smoothing*

Este método é, em parte, semelhante ao anterior mas permite adicionar parâmetros em relação à tendência e sazonalidade dos dados, o que permite um *forecast* mais correto caso os dados possuam realmente alguma regularidade tendo em conta o tempo analisado.

11 Análise preditiva

Depois de aplicados os métodos que referimos no capítulo anterior, tratámos de inferir sobre alguns resultados futuros relativamente aos imóveis que tínhamos provenientes do processo de *webscraping*. Antes de mostrarmos alguns exemplos, devemos referir que a quantidade de dados que tínhamos para fazer este processo era muito reduzida, uma vez que os dados foram inicialmente recolhidos numa altura tardia do projeto devido a várias dificuldades com o *webscraping* e que consideramos insuficientes para um, pelo menos, razoável *forecast*, no entanto, decidimos proceder à previsão mesmo tendo em conta todos os entraves a este processo.

Como exemplo, decidimos pegar mais uma vez no id do imóvel da figura 54 para se poder comparar os preços extraídos com a previsão futura. Nas imagens a seguir, iremos apresentar os resultados obtidos na utilização do método de previsão referenciados.

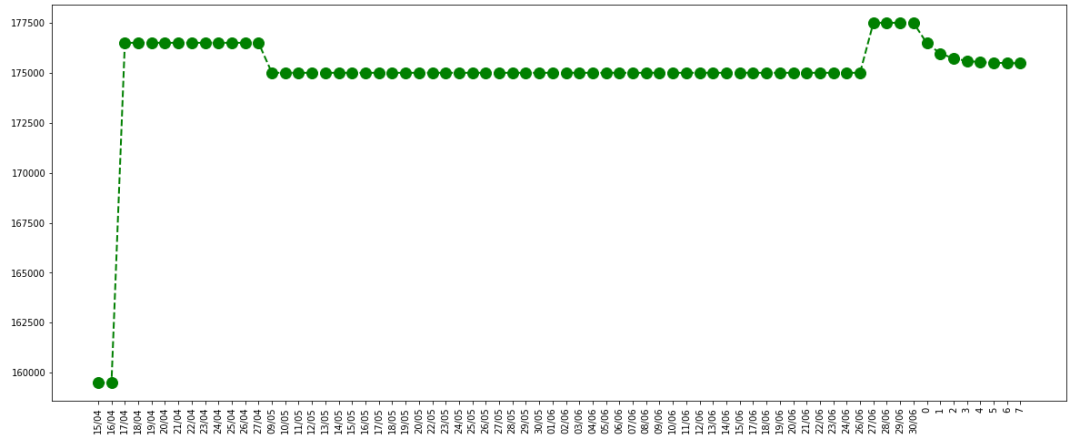


Figura 55: Gráfico com os preços e respetivo *forecast* do imóvel com o id 11160514 utilizando *Auto Regression*

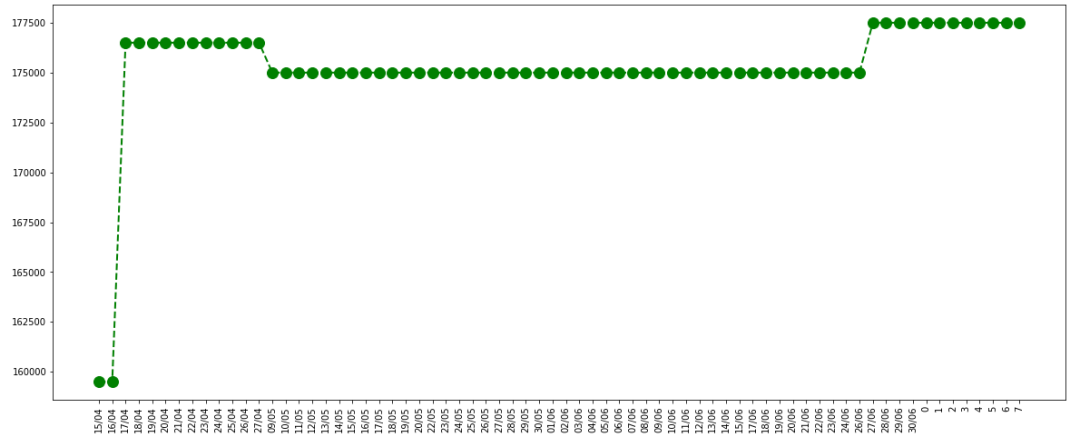


Figura 56: Gráfico com os preços e respetivo *forecast* do imóvel com o id 11160514 utilizando *Simple Exponential Smoothing*

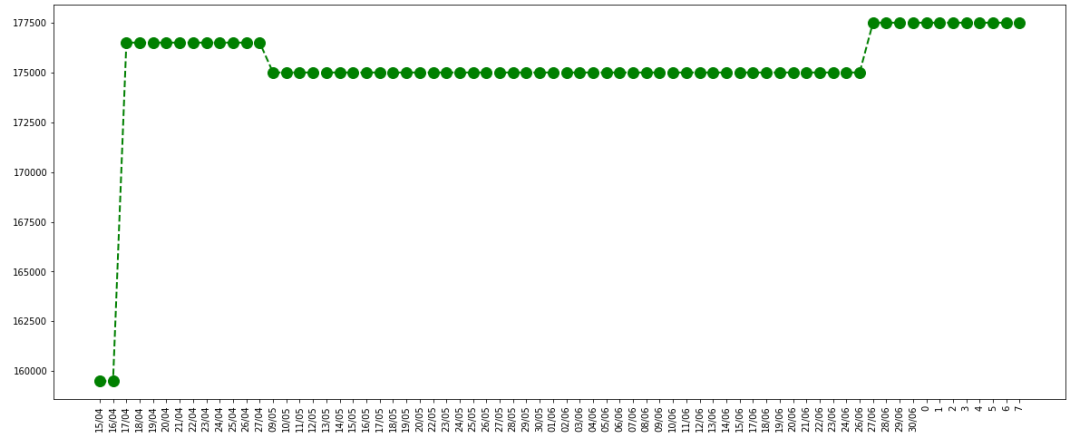


Figura 57: Gráfico com os preços e respetivo *forecast* do imóvel com o id 11160514 utilizando *Holt Winter's Exponential Smoothing*

Pelos resultados que obtivemos podemos ver que existe uma ligeira diferença entre o gráfico em que foi utilizada *Auto Regression* e os restantes 2 pois podemos ver que existe um pequeno decréscimo do mesmo seguido do que parece ser uma regularização. Nos outros 2 gráfico o que acontece é simplesmente a continuação do último preço registado sem que haja qualquer oscilação na previsão feita. De realçar que mesmo alterando um ou outro parâmetro da *Holt Winter's Exponential Smoothing* os resultados foram semelhantes.

Gostaríamos apenas de dizer, antes de concluir, que também tentamos obter resultados utilizando redes LSTM mas não havia como inferir na *performance* da rede, por isso, decidimos deixar de fora utilização deste método bem como os seus resultados.

12 Conclusão

Dado por concluído o trabalho, surge a necessidade de avaliar a solução final tendo em conta o problema apresentado na introdução. O **ScrapMyProp** atingiu o objetivo inicial de coleção de dados para o concelho de Braga, identificação de preços de imóveis e previsão da tendência dos preços.

Um das limitações com o qual o grupo se deparou foi a dificuldade em obter dados dos *websites* das imobiliárias, devido às políticas de privacidade que estes possuem para ter controlo sobre a concorrência. Esta questão afetou negativamente a qualidade da informação obtida pois não foi possível assim obter tanto uma maior quantidade de dados, assim como uma série temporal mais longa. Caso este problema de segurança não tivesse existido, esperava-se que tanto a identificação dos preços dos imóveis como a previsão a longo prazo teriam tido melhores desempenhos, dado em conta a maior variabilidade e longevidade dos dados.

Apesar do que foi realizado, ficam bastantes ideias como trabalho futuro pois este assunto, no geral, possui muita investigação por efetuar visto tratar-se dum mercado com muitas condicionantes envolvidas. Em relação aos dados utilizados neste trabalho, uma possibilidade a investigar seria testar os dados colecionados com outros valores para os limites dos *outliers* ou da percentagem de dados a retirar no tratamento de dados em falta. Outra possibilidade seria a geração sintética de dados para aumentar a quantidade de informação, que seria uma alternativa também para resolver o problema da pouca quantidade da dados devido à privacidade dos *websites* das imobiliárias. Essencialmente, caso as políticas de privacidade consigam ser acordadas, seria possível obter mais informação com diferentes fatores influenciadores do preço e, assim, incluir mais detalhe sobre todas as condicionantes nos modelos utilizados. Além dos dados de vendas foram também recolhidos dados com valores das rendas. O trabalho futuro pode também passar por essa vertente na expectativa de visualizar maiores variações nos preços, tendo em conta que os valores de rendas são mais instáveis devido à facilidade de mudança, ao contrário dos preços de venda dos imóveis.

Além dum maior conhecimento imobiliário que este projeto proporcionou a nível regulamentar, foi também uma fonte de pesquisa e aperfeiçoamento de várias técnicas, desde o *webscraping* até aos algoritmos de *machine learning*, tanto para problemas de regressão como para problemas de previsão.

Referências

[1] Ruy Figueiredo. "Manual de Avaliação Imobiliária", 7º Edição, Valor M2 LDA, 2018.

[2] Pedro Manuel Gameiro Henriques. "Avaliação Imobiliária", Departamento de Engenharia Civil - Técnico Lisboa, 2012/2013.

[3] Keller Williams. "Avaliação Imobiliária", https://ana-macao-kw.pt/avaliacao-imobiliaria.

[4] Código Civil Decreto-Lei n.º 47344/66 - 25/11. Diário da República Eletrónico, https://dre.pt/web/guest/legislacao-consolidada/-/lc/106487514/201909180701/73407280/diploma/indice.

[5] Decreto-Lei n.º 287/2003 - 12/11. Autoridade Tributária e Aduaneira, http://info.portaldasfinancas.gov.pt/pt/informacao_fiscal/codigos_tributarios/cimi/Pages/cimi2.aspx.

[6] Código das Expropriações Lei n.º 168/99 - 18/09. Diário da República Eletrónico, https://dre.pt/web/guest/legislacao-consolidada/-/lc/view?cid=436.

[7] Valor Patrimonial Tributário. "ComparaJá", https://www.comparaja.pt/blog/valor-patrimonial-tributario.

[8] Decreto-Lei n.º 287/2003 - 12/11. Autoridade Tributária e Aduaneira, http://info.portaldasfinancas.gov.pt/pt/informacao_fiscal/codigos_tributarios/cimi/Pages/codigo-do-imi-indice.aspx.

[9] Portaria n.º 156/2014. Diário da República Eletrónico, https://dre.pt/pesquisa/-/search/56053277/details/maximized.

[10] Y. Li, Q. Pan, T. Yang and L. Guo, "Reasonable price recommendation on Airbnb using Multi-Scale clustering,"2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 7038-7041.

[11] Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.

[12] Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data, Expert Systems with Applications 42(6): 2928–2934.

[13] Aminah Md Yusof and Syuhaida Ismail. Multiple Regressions in Analysing House Price Variations (2012). Universiti Teknologi Malaysia, Johor, Malaysia.

[14] Martijn Duijster. The predictive power of house price forecasting models. University of Amsterdam.

[15] Aswin Sivam Ravikumar. Real Estate Price Prediction Using Machine Learning (2018). School of Computing National College of Ireland.

[16] Neelam Shinde and Kiran Gawande. Valuation of House Prices using Predictive Techniques (2018). Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India.