Previsão do tempo de venda de um imóvel recorrendo a *data analytics*

Nadine Santos Carvalho

Dissertação de Mestrado

Orientador na FEUP: Prof. José Luís Moura Borges

Mestrado em Engenharia e Gestão Industrial

2022-07-11

À minha família.

Resumo

A EURO BROKERS – Consultoria & Serviços Imobiliários é uma empresa de consultoria e serviços imobiliários que procura trazer um conceito personalizado de soluções e atendimento de referência aos seus clientes.

A empresa foca-se em prestar um serviço de excelência no segmento imobiliário, com base na qualificação e treino dos seus profissionais, inovações tecnológicas e mercado digital. Por este motivo, surgiu a oportunidade de desenvolver um projeto de *data analytics* que representasse uma vantagem competitiva para a empresa: previsão do tempo de venda de um imóvel.

A presente Dissertação foi realizada no âmbito da Unidade Curricular Dissertação para a conclusão do Mestrado em Engenharia e Gestão Industrial da Faculdade de Engenharia da Universidade do Porto.

Numa primeira fase, foi essencial uma etapa de formação na empresa, focando-se na importância de obter uma visão geral sobre as fases necessárias na área imobiliária e, posteriormente, foi necessário recolher os dados e métricas para a execução da análise descritiva e preditiva.

Através do método *CRISP-DM*, foram desenvolvidos quatro algoritmos: árvores de regressão, redes neuronais, *nearest neighbor regression* e *random forest*, recorrendo ao *software* R. A base de dados original continha observações fora da cidade do Porto, em número insuficiente e não representativas do tempo de venda médio dos imóveis comercializados pela EURO BROKERS. Por esse motivo, foi criada uma sub-base de dados, constituída apenas por imóveis da cidade do Porto. Com o cálculo do erro raiz quadrada médio, foi possível concluir que o melhor modelo de previsão utiliza redes neuronais, com 1,8 meses de discrepância relativamente aos valores reais.

Este projeto contribuiu, assim, não só para uma análise profunda da empresa, mas também para uma possível implementação desta métrica em futuras angariações.

Forecasting the time of sale of a property using data analytics

Abstract

EURO BROKERS – Consultoria & Serviços Imobiliários is a consultancy and real estate services company, which aims at offering highly customized solutions and referency to its customer services.

The company focuses on providing an excellent service in the real estate segment, based on the qualification and training of its professionals, technological innovations and the digital market. Thus, an opportunity to develop a data analytics project arose, aiming at bringing a competitive advantage for the company, through the forecasting of the time of sale of a property.

The Dissertation was developed in the Dissertation Course Unit, for the conclusion of the Masters in Engineering and Industrial Management at the Faculty of Engineering of the University of Porto.

In a first phase, a training stage in the company was crucial, focused on the importance of obtaining an overview of the necessary phases in the real estate area. Furthermore, collecting data and metrics for the performance of the descriptive and predictive analysis was necessary.

Using the CRISP-DM method, four algorithms were developed: regression trees, neural networks, nearest neighbor regression and random forest, using the R software. The original database contained a low number observations outside the city of Porto, which were not representative of the overall average time of sale of the properties EURO BROKERS has on the market. Thus, a sub-database was created, consisting only of properties in the city of Porto. The calculation of the root mean square error enabled the conclusion that the best prediction model is the one which uses neural networks, with 1.8 months of discrepancy in relation to the real values.

Therefore, this project contributed to an in-depth analysis of the company and to an eventual implementation of this metric for future property raising efforts.

Agradecimentos

Dedico a conclusão desta etapa a todos aqueles que fizeram parte do meu percurso, que me aconselharam e apoiaram.

Um agradecimento especial ao meu orientador Thiago Mello, e ao seu sócio Ricardo Frias, bem como a toda a equipa EURO BROKERS, que me acolheram e ensinaram ao longo de todo o estágio, esclarecendo todas as minhas dúvidas e acompanhando-me de forma contínua.

Ao meu Orientador da FEUP, Professor José Luís Moura Borges por todos os conselhos e constante disponibilidade.

Ao Professor José Coelho Rodrigues pela ajuda crucial na execução do projeto de dissertação.

À minha família, que para além de me acompanharem ao longo do meu percurso académico, também me inspiraram de uma forma especial, e ajudaram em todas as dimensões.

Índice de Conteúdos

1 Introdução 1
1.1 Enquadramento do projeto na empresa 1
1.2 Apresentação da EURO BROKERS 2
1.3 Objetivos do projeto 3
1.4 Metodologia 4
1.5 Estrutura da dissertação 5

2 Enquadramento Teórico 6
2.1 Conceitos base de Data Analytics 6
2.1.1 Análise Descritiva 8
2.1.1.1 Association Rules 9
2.1.2 Análise Preditiva 10
2.1.2.1 Algoritmos 11
2.2 Big Data no ramo imobiliário: o que é, como é utilizado e onde encontrar 11
2.3 Estratégias de Data Analytics existentes no mercado imobiliário 12
2.3.1 Apoio à tomada de decisão 12
2.3.2 Índices de Preços de Propriedades 13
2.3.3 Modelos de Avaliação Automatizada 13
2.4 Envolvimento do Cliente 14

3 Descrição da situação atual 15
3.1 Processo de angariação 15
3.2 Métricas da empresa 17
3.3 Base de dados da empresa e respetivas variáveis 19
3.3.1 Clientes Proprietários 19
3.3.2 Clientes Compradores 21
3.3.3 Imóveis 24

4 Modelo e Soluções 29
4.1 Preparação dos dados 29
4.2 Modelos de Previsão 33
4.2.1 Árvores de Regressão 34
4.2.2 Redes Neuronais Artificiais 38
4.2.3 Nearest Neighbor Regressão 39
4.2.4 Random Forest 40
4.2.5 Resumo de Resultados obtidos 42
4.3 Metodologia de implementação 43

5 Conclusões e perspetivas de trabalho futuro 44

ANEXO A:Resultados das Análises 48

Siglas

CP – Complex Parameter

CRISP-DM - Cross Industry Standard Process for Data Mining

KDD - Knowledge Discovery in Databases

k-NN - K-nearest neighbors

MAE - Mean absolute error

MSE - Mean squared error

RMSE - Root Mean Square Error

Índice de Figuras

Figura 1 - Cronograma do projeto de dissertação 5
Figura 2 - Fases do método KDD (Fayyad, Piatetsky-Shapiro, and Smyth 1996) 8
Figura 3 - Boxplot e suas características ("Análise Descritiva" 2013) 10
Figura 4 - Variação de *Clicks* e *Leads* desde 2018 16
Figura 5 - Volume de Negócios de 2018 a 2022 16
Figura 6 - Percentagens de *Leads*, Reservas e Negócios Realizados 17
Figura 7 - Gráfico de Leads/Visitas/Reservas 18
Figura 8 - Número de *Leads* por Tipologia 18
Figura 9 - Quantidade de imóveis por observação (por proprietário) 19
Figura 10 – Número de Proprietários por Concelho 20
Figura 11 - Número de Proprietário por Ano 20
Figura 12 - Conjuntos de Itens Frequentes e respetivos suportes da base de dados Proprietários 21
Figura 13 - Regras para os conjuntos de itens frequentes considerados 21
Figura 14 - Tipo de Cliente 22
Figura 15 - Número de Clientes por Ano 22
Figura 16 – Regras obtidas no método *Association Rules* para a base de dados dos compradores 23
Figura 17 - Conjuntos de Itens Frequentes, e respetivos suportes após diminuição do *support* 23
Figura 18 - Regras para os conjuntos de itens frequentes considerados, e respetivas métricas 23
Figura 19 - Imóveis por Localidade 24
Figura 20 - Total de Imóveis por Tipologia 25
Figura 21 - Quantidade de Imóveis por Estado de Utilização 25
Figura 22 - *BoxPlot* e medidas relativas ao Preço 26
Figura 23 - Classe Energética em função do Estado do Imóvel 27
Figura 24 - Quantidade de Imóveis por cada Classe Energética 27
Figura 25 - Regras obtidas do método *Association Rules*, e respetivas métricas, para a base de dados Imóveis 28
Figura 26 - *Accuracy* em função do número de vizinhos 30
Figura 27 - *BoxPlots* e respetivas medidas da Área Útil e Bruta 31
Figura 28 - Matriz de Correlação de *Pearson* 32
Figura 29 - Matriz de correlação de *Cramer's V* 33
Figura 30 - Gráficos de Dispersão Comissão, Preço e Área Útil 34
Figura 31 - Possíveis CP's para Árvore de Regressão 35
Figura 32 - Árvore de Regressão dos Imóveis do Porto 37

Figura 33 - Gráfico comparativo entre valores reais e previstos 37
Figura 34 - Gráficos de Comparação entre valor real e preditivo da base de dados original e base de dados do Porto, respetivamente 39
Figura 35 - Valores de k e respetivos erros 39
Figura 36 - Comparação do RMSE com *caret* e *parameter tunning* 40
Figura 37 - Resultado do modelo *Random Forest* na base de dados original 41
Figura 38 - Aumento de MSE por variável 41
Figura 39 - Resultados do modelo *Random Forest* para base de dados do Porto 42
Figura 40 - Número de Clientes Compradores por Objetivo de negócio 48
Figura 41 – BoxPlot do Tempo de Venda da Base de Dados Original 48
Figura 42 - Resultados da Árvore de Regressão da Base de Dados Original 49
Figura 43 - Gráfico comparativo entre valores reais e previstos da base de dados original, no modelo Random Forest 50
Figura 44 - Comparação do RMSE com caret e parameter tunning e dados reais e previsto, respetivamente, da base de dados do Porto, no modelo Random Forest 50
Figura 45 - Aumento de MSE por variável da base de dados do Porto 51
Figura 46 - Variáveis após transformação em *dummy variables* 51

Índice de Tabelas

Tabela 1 - Tempo de venda dos imóveis por distrito. Retirado de Idealista.pt, acedido em abril de 2022 3

Tabela 2 - Métricas de Avaliação das Árvores de Decisão 36

Tabela 3 Métricas de Avaliação das Árvores de Decisão da Base de Dados do Porto 37

Tabela 4 - Métricas de Avaliação das Redes Neurais 38

Tabela 5 - Métricas de Avaliação do *k-NN* 40

Tabela 6 – Erros, em meses, associados aos diferentes modelos 42

1 Introdução

O sucesso de uma empresa de consultoria imobiliária está implicitamente ligado à sua capacidade de avaliar riscos e oportunidades futuras de um local ou investimento. O surgimento de soluções de *data analytics* veio revolucionar a tomada de decisão, até então baseada unicamente em intuição e dados retrospetivos tradicionais - dados estruturados, de pequena dimensão e de fácil manipulação e interpretação - como transações de vendas. Através desta tecnologia, as empresas do ramo em questão têm acesso a uma série de novas análises, nomeadamente os índices de preços de propriedades, que aumentam a sua competitividade e eficiência quando implementadas na empresa.

O objetivo de se recorrer a *real estate data analytics* prende-se com a necessidade de organizar e estudar grandes quantidades de dados de forma a, através destes, identificar e gerenciar riscos, prever o comportamento do cliente e automatizar processos (Park 2020). Geralmente, consideram-se quatro tipos de *data analytics* com base no objetivo da análise, e que podem ser aplicados a qualquer ramo, nomeadamente ao ramo imobiliário: análise descritiva, análise de diagnóstico, análise preditiva e análise prescritiva. Este processo recorre a sistemas e *softwares* especificamente desenvolvidos para investidores imobiliários.

Este projeto pretende prever o valor de uma determinada variável intrínseca à análise imobiliária – o tempo de venda de um imóvel (em meses) - assim como investigar a influência direta ou indireta de outras variáveis, como a tipologia e localização do imóvel. Para atingir este fim, torna-se indispensável recorrer a *big data* que será compilada a partir de dados, pesquisas de negócios, e recolha de informações disponíveis *online*.

Espera-se que o maior detalhe na análise dos imóveis permita à empresa aumentar a sua credibilidade e confiança junto dos seus atuais e futuros clientes proprietários, uma vez que irá fornecer um maior grau de certeza relativamente à venda do seu imóvel. Internamente, a empresa conseguirá planear/organizar de uma forma mais eficaz os seus recursos (humanos e financeiros), pois terá uma visão global de quando irá reduzir o seu portefólio, consequência das vendas dos imóveis nele presentes.

1.1 Enquadramento do projeto na empresa

O presente projeto de dissertação foi realizado no âmbito da unidade curricular de Dissertação do 5º ano do Mestrado Integrado de Gestão e Engenharia Industrial da Faculdade de Engenharia da Universidade do Porto. O projeto foi desenvolvido durante o segundo semestre do ano letivo 2021/2022, na empresa EURO BROKERS – Consultoria & Serviços Imobiliários, uma empresa de consultoria e serviços imobiliários que procura trazer um conceito personalizado de soluções e atendimento de excelência aos seus clientes.

Aquando da escolha, por parte de um cliente proprietário, de uma consultora imobiliária, este tende a dar primazia à diversidade e qualidade. A oferta significativa do mercado imobiliário, um dos mais competitivos a nível nacional e global, torna essencial uma empresa estar na

vanguarda do seu setor, oferecendo um serviço de excelência que permita que os imóveis sejam promovidos e vendidos pelo maior valor possível, da forma mais célere.

Um dos maiores desafios no decorrer da atividade de um consultor imobiliário prende-se com a angariação de clientes proprietários que necessitam de assistência na venda do seu imóvel. Para o sucesso desta operação, a realização prévia de estudos de mercado de elevada qualidade torna-se essencial.

A análise realizada na presente dissertação foca-se na previsão do número de meses – contabilizado entre a etapa de angariação e a assinatura da escritura – até à venda do imóvel, sendo esta a variável a estudar.

A difícil previsão, com rigor, desta variável, deve-se ao facto de esta ser influenciada por características intrínsecas ao imóvel, tais como tipologia, localização, ano de construção, bem como por fatores externos imprevisíveis, incluindo crises económicas, sociais, ambientais ou guerras. Como tal, de forma a endereçar os desafios mencionados, este projeto recorrerá a soluções de *data analytics*, que utilizarão o método *CRISP-DM*.

O impacto da implementação do presente projeto na EURO BROKERS prender-se-á com dois fatores essenciais. O primeiro está associado a uma melhoria significativa da qualidade de serviço oferecida aos seus clientes proprietários, devido ao conhecimento gerado pela análise de mercado realizada, que se traduzirá numa maior confiança deste na empresa. O segundo refere-se à otimização dos processos logísticos e da eficiência de recursos da empresa, devido ao conhecimento do momento de venda de cada imóvel, permitindo reforçar campanhas de marketing e investir/desinvestir em ações de angariação.

1.2 Apresentação da EURO BROKERS

A EURO BROKERS surgiu em 2018, tendo os fundadores Thiago Mello e Ricardo Frias, que identificaram uma lacuna/necessidade no mercado, nomeadamente a falta de acompanhamento ao cliente, e a competição excessiva entre colegas do ramo. Atualmente, emprega 5 pessoas no Porto e 1 no Brasil e providencia uma ampla gama de serviços, desde a identificação e qualificação do perfil dos clientes, apoio jurídico, projetos de arquitetura, intermediação de crédito habitacional, seguros, entre outros.

A empresa foca-se em prestar um serviço diferenciado no segmento imobiliário, com base na qualificação e treino dos seus profissionais, inovações tecnológicas e mercado digital, respeitando os mais elevados padrões de ética, discrição e integridade, conseguindo assim, um crescimento sustentável da marca.

Atualmente, a empresa utiliza apenas soluções de *Customer Relationship Management* (*CRM*) e *Enterprise Resource Planning* (*ERP*), que não integram análises preditivas. O primeiro armazena todos os dados de clientes, incluindo as suas preferências, bem como o registo das suas interações com a empresa; o segundo apoia a atividade de gestão dos processos internos da empresa de forma integrada. Em concreto, a EURO BROKERS utiliza o sistema X-IMO como CRM, permitindo fazer uma correspondência adequada, com um elevado grau de precisão, dos imóveis às necessidades dos clientes, mas também efetuar um acompanhamento diferenciado aos clientes proprietários.

Para além deste, a empresa recorre ao *software* CASAFARI, uma base de dados com milhões de propriedades onde são realizadas análises comparativas de mercado, impulsionadas por inteligência artificial (IA) e contactos de particulares.

Importa referir que, até ao momento, a empresa não dispõe de nenhum *software* que permita prever o tempo de venda de um imóvel. É apenas conhecida a média do tempo de venda baseada em informações disponíveis online, nomeadamente as referidas na Tabela 1. A média geral apresentada, discriminada por distrito, compara dois períodos temporais, sem ter em

conta qualquer outro fator que não o geográfico. Como tal, torna-se impossível determinar, com base nestes valores, que outras variáveis poderiam impactar o tempo de venda do imóvel, bem como a extensão dessa influência. Adicionalmente, esta análise baseia-se no tempo decorrido entre a entrada do anúncio numa plataforma online e a retirada deste, não contabilizando o período entre a etapa de angariação e a colocação do anúncio, nem confirmando o momento da assinatura da escritura. Embora seja possível retirar informações desta análise, estas terão pouco valor para o proprietário, devido à elevada incerteza associada.

A EURO BROKERS, na sua procura constante de introduzir melhorias nos seus processos internos, acrescentando valor aos seus clientes, pretende introduzir um método de análise rigoroso e multivariável na sua atividade.

Tabela 1 - Tempo de venda dos imóveis por distrito. Retirado de Idealista.pt, acedido em abril de 2022

Distritos	3T2020 (meses)	3T2019 (meses)	Variação (dias)
Portugal	**5,1**	**4,3**	**22**
Aveiro	5,0	4,9	**4**
Beja	8,1	7,4	**19**
Braga	4,9	4,8	**2**
Bragança	7,0	6,9	**4**
Castelo Branco	7,0	6,2	**25**
Coimbra	6,0	5,5	**13**
Évora	5,1	5,9	**-24**
Faro	6,4	4,7	**50**
Guarda	6,1	8,3	**-67**
Leiria	6,5	5,3	**37**
Lisboa	4,1	3,8	**10**
Portalegre	7,7	7,9	**-5**
Porto	4,4	3,7	**20**
Santarém	6,2	5,8	**14**
Setúbal	3,6	3,8	**-3**
Viana do Castelo	8,3	6,2	**61**
Vila Real	8,2	5,5	**81**
Viseu	6,3	5,3	**29**
Madeira (Ilha)	6,7	5,1	**48**
São Miguel (ilha)	4,1	5,8	**-52**

1.3 Objetivos do projeto

O presente projeto pretende colmatar uma necessidade existente na EURO BROKERS, que se deve aos métodos pouco precisos presentemente utilizados pela empresa para prever o tempo de venda de um imóvel. Este objetivo geral pressupõe um estudo rigoroso das diferentes estratégias de *data analytics*, aplicáveis à realidade da empresa. Para a sua prossecução, serão necessários, portanto:

- A sistematização das dificuldades que foram encontradas até à data na implementação de soluções de *data analytics* na empresa;
- O estudo do potencial das soluções de *data analytics* no ramo imobiliário (*market research*) e respetivos competidores (*market analysis*);

- A recolha de dados e respetiva análise recorrendo ao *software* R, que será constituída por duas etapas fundamentais:
    - *Descriptive data analysis*, recorrendo a estatística descritiva, onde o objetivo será identificar grupos maioritários, perceber qual o número de clientes, leads, qual a tipologia mais frequente, de forma a obter uma visão geral da empresa e as suas métricas;
    - *Predictive data analysis*, cujo objetivo será prever o tempo de venda de um imóvel, em meses.
- A comparação das diferentes soluções a aplicar pela empresa, incluindo a metodologia a seguir, e potenciais impactos que estas terão.

1.4 Metodologia

Após um período inicial de formação nas instalações da EURO BROKERS, que permitiu um levantamento das principais limitações na atividade da empresa, sobretudo no que concerne à abordagem simplista relativa à determinação de variáveis associadas à venda de imóveis, foram definidos os principais objetivos na elaboração do presente projeto (elencados na secção anterior).

A etapa de formação compreendeu a execução e acompanhamento de tarefas essenciais da empresa em estrita colaboração com os seus restantes elementos, nomeadamente a organização de visitas, contacto com o cliente, prospeção na cidade e análise concorrencial. Adicionalmente, foi necessária a familiarização com os *softwares* utilizados na empresa, nomeadamente o sistema X-IMO, para a determinação das limitações e dificuldades em implementar uma nova metodologia de suporte à atividade da EURO BROKERS.

Para atingir os objetivos identificados, foi necessário, paralelamente, realizar um estudo aprofundado dos diferentes métodos associados à *data analytics*, bem como da sua aplicabilidade ao segmento imobiliário, especificamente à realidade da empresa. Esta tarefa permitiu identificar a necessidade de uma recolha de dados que foram posteriormente tratados para a previsão e análise descritiva. A base de dados correspondente aos imóveis da empresa foi tratada com *predictive data analysis*, realizada através do método *CRISP-DM*, com as fases de construção:

- seleção dos dados;
- pré-processamento destes;
- transformação;
- modelação; e
- avaliação e desenvolvimento.

A limpeza e normalização de dados são processos imprescindíveis para aumentar a precisão do modelo analítico desenvolvido, sendo estas etapas, inseridas na seleção e pré-processamento de dados, as mais importantes e morosas.

Os modelos construídos dizem respeito ao *k-NN*, *Neural Networks*, Árvore de Regressão e finalmente, *Random Forest*. A avaliação de cada método será realizada através do cálculo de erros inerentes à previsão da variável tempo de venda, nomeadamente o erro raiz quadrada médio.

Em relação aos clientes proprietários e compradores, a análise foi, maioritariamente, do tipo descritiva, recorrendo a estatísticas descritivas, como média, desvio padrão e *boxplots*, mas também recorrendo ao método *Association Rules*.

Tendo em conta o exposto, foi proposta uma metodologia adequada, de forma a permitir alcançar os objetivos definidos. A Figura 1 trata-se de uma representação gráfica do cronograma que descreve todas as etapas do projeto.

Figura 1 - Cronograma do projeto de dissertação

1.5 Estrutura da dissertação

A dissertação está dividida em cinco capítulos.

O primeiro capítulo descreve o tema do projeto, assim como o seu enquadramento na empresa EURO BROKERS, enumerando os principais objetivos a alcançar e a metodologia proposta.

No segundo capítulo é apresentado o enquadramento teórico de vários conceitos de *data analytics* que são explorados ao longo do projeto e que servirão de base à posterior análise do mercado imobiliário, mais concretamente da EURO BROKERS. Ainda neste capítulo é identificado o potencial de aplicação de *big data* no setor imobiliário em geral e de que forma são atualmente utilizadas as estratégias de *data analytics* existentes.

No terceiro capítulo é identificada a situação atual da empresa, através do levantamento de informações importantes ao seu funcionamento, destacando-se o processo de angariação, métricas da empresa e base de dados da empresa e respetivas variáveis.

No quarto capítulo apresenta-se a forma como a base de dados é pré-processada, e são construídos os modelos de previsão e identificados os respetivos erros. É salientado o modelo mais eficaz e de que forma pode ser utilizado posteriormente.

No último capítulo são apresentadas as conclusões do projeto desenvolvido, assim como algumas sugestões de possível investigação futura para complementar/otimizar/aprofundar o tema estudado.

2 Enquadramento Teórico

Neste capítulo são abordados os variados conceitos que o projeto abrange. Os temas em análise são a *data analytics* e suas bases, *big data* no ramo imobiliário, estratégias de *data analytics* já existentes no mercado e a forma como todas estas dimensões influenciam o envolvimento do cliente com a empresa.

2.1 Conceitos base de Data Analytics

A *data analytics*, com objetivos e definições específicos, envolve um amplo campo de técnicas, para encontrar padrões e tendências a partir do processo de análise de *raw data*.

Dos primórdios da *data analytics* destacam-se dois estudos: o primeiro censo populacional conhecido, realizado pelo governo sueco em 1749, assim como o estudo académico britânico de Richard Doll sobre o cancro nos pulmões e o tabaco em 1950. Desde então, a análise de dados tem sido responsável pela estimulação de produção de dados na procura de conhecimento. Cada um destes estudos procurou encontrar respostas para melhor compreender e clarificar diferentes temas. Efetivamente, o potencial de aplicabilidade da *data analytics* é incomensurável, uma vez que a sua utilização permite esclarecer e aferir conhecimento sobre as mais diversas áreas.

A *data analytics* é uma ciência que examina dados brutos com o intuito de encontrar tendências e padrões, de forma a retirar conclusões sobre essa informação, aplicando um processo algorítmico. Qualquer tipo de informação não processada pode ser analisada de forma a obter *insights*, utilizados posteriormente para otimizar e aumentar a eficiência de processos, ajudar na tomada de decisão e na obtenção de melhores resultados (Cattaneo et al. 2018).

Segundo a Forbes, o mercado de análise de dados está a crescer e continuará nesta tendência. De acordo com analistas da *IDC* (*International Data Corporation*), estima-se que as empresas gastaram cerca de $215B em 2021 em soluções de *big data* e análise de negócios, representando um aumento de 10% relativamente ao ano anterior (Rohit Amarnath 2022).

A *data analytics* requer um pensamento crítico estruturado, que conduzirá a um aumento de intuição e criatividade e aumentará o conhecimento específico na área. Para além do desenvolvimento e utilização de diferentes modelos, é também necessário avaliar a veracidade, lógica e coerência dos resultados obtidos, e a existência de falhas. Assim, será obtida uma estrutura e princípios que permitirão analisar sistematicamente este tipo de problemas (Runkler 2012).

Atualmente, existem quatro tipos predominantes de *data analytics*:

- Análise descritiva – Baseia-se em modelos que ajudam a compreender e dar resposta relativamente ao que aconteceu durante um determinado período de tempo;

- Análise diagnóstica – É uma forma de análise avançada que examina os dados de forma a definir o porquê de determinado evento ter acontecido. Caracteriza-se por técnicas como *drill-down*, descoberta e mineração de dados e correlações;
- Análise preditiva – Diz respeito a modelos que combinam dados históricos existentes com algoritmos preditivos, para assim determinar a probabilidade de um determinado evento acontecer no futuro. Esta análise utiliza as descobertas da análise descritiva e diagnóstica para identificar grupos e casos especiais para fazer a previsão; e
- Análise prescritiva – Fornece informações sobre decisões ideais com base nos cenários futuros previstos. Desta forma, o objetivo é eliminar um problema futuro ou aproveitar uma tendência promissora (Tsai et al. 2015).

Geralmente, as principais etapas do processo de *data analytics* dizem respeito à mineração, gestão, análise estatística e apresentação dos dados. Estas etapas adquirem diferentes importâncias dependendo dos dados a utilizar e do objetivo da análise, e são complementadas de acordo com o método utilizado.

A mineração de dados corresponde à extração de dados de fontes de dados. As principais fases da mineração de dados dizem respeito à sua extração, transformação e análise. Desta forma, os dados brutos são convertidos num formato mais intuitivo e utilizável. A análise estatística é realizada através da criação de modelos estatísticos, recorrendo a linguagens de programação. Neste projeto, a linguagem utilizada foi o R. Na etapa final, correspondente à apresentação de dados, é extremamente importante dissecar toda a história dos dados, i.e., para que os gestores consigam visualizar e compreender de forma clara os *insights* de toda a análise (Cattaneo et al. 2018).

Neste sentido, o método mais utilizado e que serve de base para a maioria dos processos de data analytics nas empresas é o *Cross Industry Standard Process for Data Mining (CRISP-DM*), que conta com seis fases distintas (Vijay Kotu 2014).

- Conhecimento do negócio – Identificação do problema que é necessário resolver. Nesta fase, são esperados três entregáveis por parte do analista à empresa: a explicação do problema e de como o projeto vai solucionar o mesmo; o objetivo principal do projeto e a definição, com clareza, da métrica a utilizar para a avaliação do sucesso do projeto;
- Compreensão dos dados – Perceção dos dados necessários e sua aquisição, e, através do recurso a estatística, exploração da qualidade dos mesmos;
- Preparação dos dados – Tratamento dos dados para modelação. Esta fase consiste em quatro tarefas principais:
    - Seleção dos dados;
    - Limpeza dos dados;
    - Construção, se necessário, de novas variáveis;
    - Integração de dados, caso existam duas bases de dados distintas.
- Modelação – Seleção do modelo mais indicado para solucionar o problema, incluindo *Regression Trees* ou *Neural Networks* (definidas no Capítulo 4). Neste processo, será necessário criar diferentes modelos, compará-los, utilizando métricas de avaliação definidas, para avaliação do mais adequado.
- Avaliação – Escolha do modelo que melhor responde aos objetivos do negócio
- Implementação – Utilização do projeto por parte da empresa, compreendendo etapas posteriores de monitorização de resultados e adaptação do modelo, se necessário.

Por outro lado, quando o objetivo se prende com a análise aprofundada dos dados e a descoberta de conhecimento através deles, sem passar pela geração de modelos e análise de negócios, é utilizado o método *Knowledge Discovery in Databases* (KDD). Este método conta com as fases descritas na Figura 2.

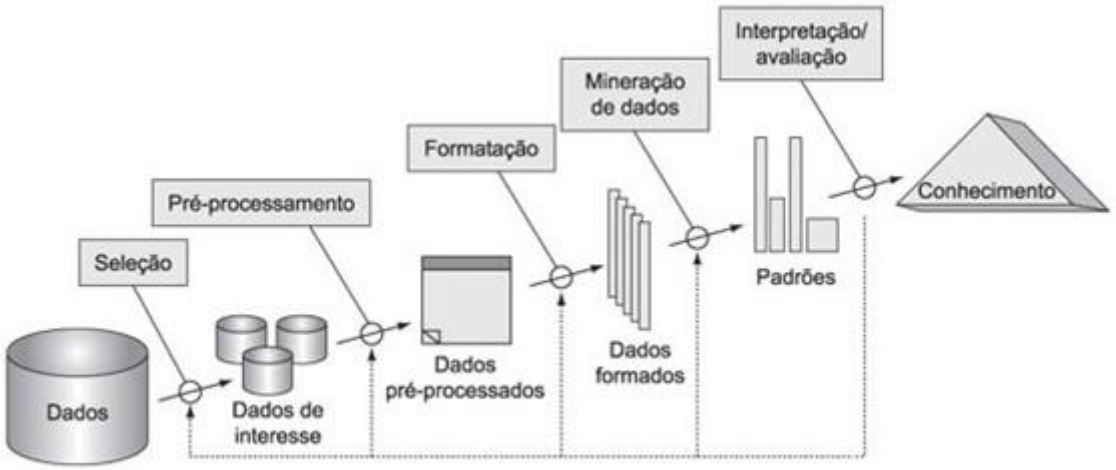


Figura 2 - Fases do método KDD (Fayyad, Piatetsky-Shapiro, and Smyth 1996)

2.1.1 Análise Descritiva

Tal como referido anteriormente, a análise descritiva diz respeito à fase inicial do estudo, permitindo a compreensão em tempo real dos acontecimentos. Nesta etapa, todos os dados são resumidos, organizados e explicados através de estatística. Normalmente, os dados são apresentados em tabelas e gráficos, podendo-se, também, recorrer a percentagens, médias e índices.

A descrição dos dados visa identificar de forma rápida e eficaz anomalias presentes nos negócios, seja por falhas e registos incorretos, ou por dados fora do padrão. Nesta análise, são também assinaladas as variáveis que contribuem para a qualidade da tomada de decisão, e podem ser de dois tipos: qualitativas (sem ordem lógica) ou quantitativas (valores numéricos).

Assim, esta dimensão possui um papel de extrema importância na *data analytics*, pois serve de suporte para análises seguintes, e permite um diagnóstico mais confiável, que se traduz em criação de estratégias e soluções robustas num plano de negócios (Reis 2002).

As métricas estatísticas mais utilizadas são as medidas de resumo numérico, como a tendência central (média, mediana e moda), dispersão (amplitude total, desvio médio absoluto, variância, desvio padrão e coeficiente de variação) e *boxplots*. De salientar que estas métricas, por si só, não são suficientes para a análise descritiva. Muitas vezes, são necessários cálculos adicionais, de forma a retirar o máximo de informações dos dados.

Um dos métodos utilizados neste contexto é o *Association Rules*, que será implementado neste projeto. Também o *Clustering* e *Anomaly Detection* são relevantes no que concerne à análise dos dados, pois o *Clustering* divide a população em grupos cujas observações são semelhantes, e o *Anomaly Detection* identifica pontos na base de dados desviados do comportamento normal.

2.1.1.1 Association Rules

Association Rules é um dos tópicos com mais relevância na mineração de dados e descoberta de conhecimento, que identifica relações entre conjuntos de itens em bases de dados. Este método permite também encontrar comportamentos associativos, de correlação e de coocorrência entre dados.

O resultado do modelo de uma análise de associação pode ser representado como um conjunto de regras do tipo: {Item A} → {Item B}. Esta regra indica, com base no histórico de todos os registos que se o Item A se verificar, há uma forte propensão de ocorrência do Item B, dentro do mesmo registo. Desta forma, o Item A é o antecedente, que pode não ser único, e o B o consequente.

De salientar, que não podem existir itens iguais no conjunto de itens antecedente e no consequente, pois existe uma relação de causalidade entre ambos.

Para encontrar as *Association Rules* recorrendo à mineração de dados, são necessárias três etapas sequenciais:

1. Preparação dos dados num formato específico, onde são criadas *dummy variables*;
2. Realização de uma lista de conjuntos de itens que ocorrem com frequência. O algoritmo de associação limita a análise aos itens que ocorrem com mais frequência, e por esse motivo o conjunto de regras final extraído na etapa seguinte é mais significativo;
3. Geração de regras de associação relevantes a partir de conjuntos de itens.

A força de uma regra de associação é, normalmente, quantificada pelo *support* e *confidence*. Estas medidas têm por base a frequência relativa de ocorrências de um determinado item definido no conjunto de dados para *training*.

O *support* de uma regra é a medida que quantifica de que forma os itens de uma regra são representados nos registos gerais. É esta métrica que indica se a regra vale a pena considerar. S*upport* elevado de uma regra, significa que favorece os itens de elevada ocorrência, e consequentemente, descobre padrões relevantes para investigação. Num outro prisma, se o *support* for baixo, significa que os itens ocorrem com pouca frequência e o relacionamento entre estes pode ser apenas por acaso.

De forma a evitar um baixo *support*, é especificado anteriormente um limite de suporte. Qualquer regra que exceda esse valor é considerada para a análise seguinte.

A *confidence* ou confiança de uma regra mede, por sua vez, a probabilidade de ocorrência do consequente da regra, de todos os registos que contem o antecedente da regra. Esta medida é calculada através da equação seguinte:

$$\text{Confidence}\,(X \text{ -> } Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}\,(X)} \tag{2.1}$$

Onde:

X, é o antecedente de regra
Y, é o consequente de regra

Através da confiança de uma regra, a frequência de ocorrência do consequente de regra é ignorada. Este facto pode levar a uma utilização de regras inadequadas, onde o consequente de regra é infrequente. De forma a ultrapassar este obstáculo, foi introduzida uma nova métrica – *lift* – um rácio que admite a independência entre o antecedente e o consequente.

O *lift* pode ser calculado da seguinte forma:

$$\text{Lift}\,(\text{X}\text{->}\text{Y}) = \frac{Support\,(X \cup Y)}{Support\,(X) * Support\,(Y)} \tag{2.2}$$

Onde:

X, é o antecedente de regra
Y, é o consequente de regra

Quanto maior o valor do *lift*, mais interessantes se tornam as regras, e mais relação as variáveis têm entre si.

O processo geral de criar regras significativas pode ser dividido em duas tarefas:

1. Encontrar todos os conjuntos de itens frequentes: Para uma análise com *n* itens, é possível encontrar 2^n – 1 conjuntos de itens frequentes (aumento exponencial de acordo com o aumento de itens).
    a. Nesta fase, é crucial definir um *support* mínimo, com o intuito de descartar conjuntos de itens menos frequentes
2. Extrair regras dos conjuntos de itens frequentes: Para o conjunto de dados com *n* itens, é possível encontrar 3^n – 2^{n+1} + 1 regras. Esta etapa extrai todas as regas com confiança superior a uma confiança mínima especificada.

O algoritmo mais utilizado para encontrar os conjuntos de itens frequentes relevantes, i.e, a cima do suporte mínimo, é o *Apriori*. *FP-Growth* também é bastante utilizado, mas não será aprofundado neste projeto.

O algoritmo *Apriori* utiliza alguns princípios lógicos para reduzir o número de conjuntos de itens a serem testados pelo *support.* Um dos princípios afirma que se um conjunto de itens é frequente, então todos os seus subconjuntos serão também frequentes, ou seja, o suporte para o conjunto de itens é superior ao mínimo imposto.

O mesmo acontece para os conjuntos de itens infrequentes, que podem posteriormente ser eliminados, caso o seu suporte seja inferior ao mínimo considerado. Assim, numa segunda fase, onde são realizadas as possíveis gerações de conjuntos de dois itens, o número destas diminuiu consideravelmente, pois foi eliminado um conjunto irrelevante.

Para a escolha de regras significativas, é então utilizada a medida de confiança. De salientar, que as regras devem ser avaliadas, posteriormente, quanto à sua validade racional, de forma a determinar se a relação é realmente útil. (Vijay Kotu 2014).

Uma forma de melhorar o modelo *Apriori* é reduzir o número de registos na base de dados.

2.1.2 Análise Preditiva

A análise preditiva tem vindo a ganhar relevância ao longo dos anos. Estima-se que foi inicialmente utilizada nas primeiras missões lunares por cientistas e engenheiros, aplicada a conceitos físico-químicos. A análise preditiva utilizada em *business analytics* baseia-se em conhecimento empírico, mais precisamente em dados históricos.

Esta análise pode ser utilizada para diversos objetivos, tais como segmentação de clientes, deteção de fraudes, análise de risco e previsão de vendas ou tempo de venda.

Neste contexto, o processo de construção do modelo preditivo ajusta-se aos dados observados, e tem duas dimensões cujos *outputs* são distintos. Por um lado, prevê o valor da variável de saída, *target*, com base nas variáveis de entrada, e, por outro lado, permite entender a relação entre estas variáveis, nomeadamente, quais das variáveis de entrada tem mais impacto na variável de saída.

Para a realização destes modelos, pode-se recorrer a diferentes tipos de algoritmos, dependendo do objetivo da análise e dos dados a analisar (Vijay Kotu 2014). Os modelos utilizados neste projeto serão aprofundados no Capítulo 4.

2.1.2.1 Algoritmos

Algoritmos são procedimentos iterativos que transformam os *inputs* em *outputs* e automatizam o processo de obtenção da melhor solução para o problema. A aplicação destes algoritmos sofisticados para extrair padrões dos dados diferencia o *data mining* das técnicas tradicionais de análise de dados.

O *data mining* define-se pela utilização de algoritmos específicos, tais como árvores de decisão, redes neuronais, regressão, entre outros. Estas técnicas, que podem ser de classificação, regressão (as mais comuns, diferindo apenas na variável *target*), associação, deteção de anomalias, série de tempo ou *text mining*, são utilizadas de acordo com o tipo de problema. Na classificação, a variável *target* adquire valor categórico ou polinomial de 0 ou 1, sim ou não, enquanto que nos problemas de regressão, o valor a prever é numérico e encontra-se dentro de um intervalo de valores.

De uma forma geral, e não pormenorizando o carácter da variável de saída, os problemas de *data analytics* podem ser categorizados em modelos supervisionados ou não supervisionados. As técnicas supervisionadas têm como objetivo prever o valor das variáveis de saída com base num conjunto de variáveis de entrada. Assim, o modelo é construído a partir de um conjunto de dados de *training* onde todas as variáveis são conhecidas.

No caso de modelos não supervisionados, é necessário realizar, a titulo de exemplo, o *clustering*, processo de identificar os grupos naturais no conjunto de dados, normalmente utilizado para segmentação de mercado (Vijay Kotu 2014).

Após a execução de vários algoritmos, é necessário avaliar o desempenho dos mesmos e compará-los, de forma a tomar uma decisão sobre qual utilizar.

2.2 Big Data no ramo imobiliário: o que é, como é utilizado e onde encontrar

O *big data* define-se, de um modo geral, como informação de elevado volume, velocidade e variedade, que, não sendo possível ser processada através de métodos tradicionais, exige técnicas inovadoras de processamento de informações. Após processamento, o *big data* permite *insights* aprimorados, tomadas de decisão mais corretas e automação de processos (Tsai et al. 2015).

Tradicionalmente, os dados imobiliários eram apenas de três tipos: financeiros, transacionais e físicos. Os dados financeiros incluem informações sobre as ações relacionadas com os imóveis e investimentos. Dados transacionais referem-se a hipotecas, compra e venda de imóveis, arrendamento, impostos, entre outros. Finalmente, os dados físicos incluem informação sobre terrenos, características do imóvel e dados de localização.

Com a introdução dos Sistemas de Informação Geográfica (SIG), novas formas de dados começaram a ser utilizadas de maneira mais eficiente no setor imobiliário. Este sistema de informação permitiu recolher dados espaciais que têm um papel de extrema importância neste setor, principalmente no que diz respeito a empreendimentos imobiliários e investimentos.

De acordo com *Winson-Geideman* e *Krause* (2016), os dados passaram a ser categorizados em três tipos principais: núcleo, espacial estático e periférico. Compreender e categorizar os diferentes tipos de dados permite aferir quais os dados dos imóveis e de que forma se devem analisar.

Os dados de núcleo representam os dados tradicionais, tais como financeiros, transacionais e físicos de cada imóvel – valor de venda, taxas a pagar, entre outros. Os de origem espacial estático dizem respeito a todos os dados fora dos limites físicos da propriedade (i.e., a vizinhança). Os periféricos não estão diretamente ligados às propriedades, mas sim à atividade das pessoas nessa localização, sendo recolhidos de fontes indiretas diferentes – procura de imóveis na internet, dados de geolocalização, informação de transito, entre outras.

2.3 Estratégias de Data Analytics existentes no mercado imobiliário

As estratégias mais utilizadas no mercado imobiliário estão focadas no apoio à tomada de decisão, cálculo de índices de preços de propriedades, modelos de avaliação automatizada, *forecasting*, *clustering*, e SIG.

Existem, atualmente, várias empresas que fornecem serviços de *data analytics* de elevada relevância multissetorial. As mais relevantes para o setor imobiliário são as seguintes:

- *Property Technology* – Utiliza IA para processar dados geoespaciais de forma a conseguir *insights* sobre compra e venda, construção, arrendamento, mudanças e investimento;
- *Cherre* – Permite que os clientes avaliem tendências e oportunidades de imóveis comerciais com elevada rapidez. Apresenta uma visão de todo o portefólio por classe de ativo, geografia, e procura oportunidades de investimento fora do mercado;
- *Stratodem* – Captura dados demográficos para prever perspetivas de crescimento. Estes alimentam a tomada de decisão para as principais organizações de desenvolvimento e investimento imobiliário;
- *Placense* – Fornece dados detalhados em tempo real sobre os fluxos de clientes num determinado local. Desta forma, é possível priorizar os investimentos em áreas de alto potencial;
- *Jupiter Intelligence* – Ajuda a prever e gerir riscos ambientais (alterações climáticas, elevação do nível do mar, intensificação de tempestades e aumento das temperaturas). Permite aos consultores planear, construir e gerir ativos, e incorporar o risco climático nas avaliações de mercado (“The Top Real Estate Analytics Companies Right Now” 2021).

2.3.1 Apoio à tomada de decisão

No ramo da consultoria imobiliária, a análise de um ativo imobiliário pressupõe a utilização de dados relativos ao imóvel em consideração. A tomada de decisão, por parte dos investidores, na compra e venda de ativos depende da qualidade da informação fornecida pelas empresas de consultoria imobiliária.

Neste contexto, dados indicativos da oferta/procura, taxas de desconto, demografia, desemprego, entre outros, são dados imprescindíveis para a análise em questão, sendo progressivamente mais fiáveis, quanto maior o número de amostras.

Para a tomada de decisão final em relação a um empreendimento ou investimento, o valor atual ou futuro depende de obras, regularizações, tomadas de posse. São estas as dimensões que servem de *inputs* para os modelos de avaliação. Com base nas informações referidas anteriormente, torna-se possível realizar um enquadramento socioeconómico, fundamental para o setor imobiliário (Winson-Geideman and Krause 2016).

2.3.2 Índices de Preços de Propriedades

O preço de uma propriedade não depende apenas das suas características, visto que duas propriedades com as mesmas características (composição, exposição, localização) podem apresentar variações no preço. Este fator introduz complexidade adicional no setor imobiliário, dificultando a análise de grandes conjuntos de dados para entender o desempenho individual deste mercado.

O cálculo de médias simples das transações históricas pode conter *outliers*, por não ser possível determinar quais as propriedades a incluir ou excluir. Adicionalmente, os preços são sensíveis aos períodos do ano, o que torna as médias pouco fiáveis. Torna-se necessário encontrar soluções para este problema.

Uma das soluções encontradas foi a utilização da regressão hedónica (conceito que descreve a felicidade imputável a cada característica), e que estima a influência que vários fatores, separadamente, têm sobre o preço de um imóvel (índices de preços de propriedades). Nesta regressão, a variável dependente é o preço, e as independentes os atributos do imóvel a analisar, onde os coeficientes estimados resultantes podem ser interpretados como a importância que os compradores colocam nas várias qualidades de um imóvel (Rodrigues 2008)

Alternativamente, a análise pode ser realizada comparando apenas as mudanças de preço em propriedades que são vendidas mais que uma vez. Este processo de vendas repetidas rastreia a mudança de preços num mesmo imóvel ao longo do tempo. Os índices *US Case-Shiller* são um exemplo desta técnica (Park 2020).

Os vendedores, de um modo geral, pretendem saber de que forma pequenas ou grandes obras, como uma remodelação de uma cozinha, influenciam o valor do seu imóvel. Com os índices de preços de propriedades é possível identificar que tipo de fatores os compradores valorizam mais.

Estes métodos de indexação dos preços das propriedades permitem às imobiliárias utilizar os dados para aferir o desempenho do mercado imobiliário. Atualmente, o *big data* na sua forma bruta pode ser combinado com informações extra, tais como localidade, características da propriedade, e dados demográficos para identificar retornos de propriedades em códigos postais específicos.

2.3.3 Modelos de Avaliação Automatizada

Os modelos de avaliação automatizada têm como objetivo utilizar os dados para produzir uma estimativa de valor de mercado de uma propriedade. Este método facilita a negociação entre comprador e vendedor, nomeadamente em comunicações à distância, uma vez que é indicado um preço inicial, de forma automatizada, à propriedade.

Estas técnicas são úteis na avaliação de hipotecas e empréstimos, na medida em que reduzem, quando bem executadas, o erro do avaliador. Assim, é possível entender melhor o mercado imobiliário presente, e avaliar um preço de transação justo para um negócio.

Uma aplicação interessante desta tecnologia é vista em empresas como *Opendoor* e *Properly*, que licitam automaticamente as casas, proporcionando aos proprietários liquidez imediata para os seus ativos (Park 2020).

A *Brognoli*, empresa de inovação do ramo imobiliário, lançou uma calculadora que considera os valores médios de uma região, bem como os imóveis presentes no seu portefólio, para calcular o valor ideal de cada propriedade. Com uma análise mais aprofundada, é possível

identificar a rentabilidade que um imóvel irá produzir em 5 ou 10 anos (Cheryshenko and Pomernyuk 2021).

O preço de um imóvel pode influenciar o tempo de venda de um imóvel, pois um preço acima do mercado aumenta o seu tempo de venda, e um preço abaixo do ideal, acelera o processo de venda, de um modo geral.

2.4 Envolvimento do Cliente

Com o aumento da informação disponível e um acesso mais facilitado à mesma, os consumidores realizam uma pesquisa e análise muito mais longa e aprofundada antes de efetuar uma compra ou venda de um imóvel.

A disseminação de ferramentas de pesquisa de imóveis cria condições para que os potenciais clientes tenham a perceção de dominar o processo de compra e venda, sentindo que não necessitam de consultores imobiliários para o sucesso de uma transação imobiliária. Assim, e de acordo com Philip Kotler, "conquistar um novo cliente custa entre 5 a 7 vezes mais do que manter um atual".

No ramo imobiliário, a retenção de clientes tem uma especial importância, visto que há um *cash-flow* constante em todas as transações. O cliente proprietário que vendeu o imóvel, ficou agora com capital para investir noutro, e assim sucessivamente.

Assim, torna-se essencial analisar a estratégia de *customer success*: estratégia focada em reter clientes por mais tempo e consequentemente reter mais receita. Traduz uma área que proporciona à empresa um diferencial, onde responde aos desafios e necessidades do cliente, antecipando os mesmos (Damin 2019).

Numa primeira fase, o mais importante é focar nos clientes de maior potencial, ou seja, aqueles com maior probabilidade de sucesso, tendo em conta o serviço que a empresa oferece. Torna-se necessário compreender qual a forma de comunicação a utilizar para captação dos mesmos.

Após conclusão da transação, o consultor imobiliário deve ainda manter o acompanhamento do cliente, para que este continue a recorrer aos serviços da empresa no futuro e contribuir para a captação de novos clientes.

3 Descrição da situação atual

Neste capítulo, analisam-se os principais elementos que fazem parte do projeto. Inicia-se a análise apresentando o processo de angariação e as dificuldades da empresa, nomeadamente na situação pandémica, e posteriormente de que forma estas se refletem no desempenho geral e específico da mesma.

De seguida, é apresentada a recolha de dados efetuada e descrita cada uma das variáveis, tendo em consideração a sua importância para a análise.

Por fim, menciona-se de que forma se vai melhorar a interação da empresa com as soluções de *data analytics* e quais os objetivos iniciais do projeto.

3.1 Processo de angariação

O processo de angariação numa empresa de consultoria imobiliária é um dos processos mais desafiantes e exigentes, e que, como tal, requer a inclusão de estratégias de *data analytics* nas empresas.

A base do processo de angariação prende-se com o facto de ser necessário manter uma carteira de imóveis diversificada para corresponder às expetativas dos diversos clientes. Neste sentido, é necessário que o consultor execute de forma organizada um plano de angariação que, normalmente, tem por base a análise das melhores localizações, vistas e ambiente externo.

A procura de imóveis pode ser reativa ou ativa. Na primeira, o consultor é contactado por um cliente à procura de um imóvel com características especificas, realiza uma pesquisa de mercado, entende a necessidade do cliente, e, caso não disponha do imóvel pretendido, realiza parcerias com outras empresas de mediação imobiliária. Na segunda, o consultor tem a necessidade de conhecer bem as diferentes localidades, de forma a escolher qual a zona de captação com maior potencial. Com o avanço da tecnologia e redes sociais, a angariação também pode ser realizada através de fóruns, blogs e portais de anúncios, em que a taxa de conversão tende a ser superior.

O processo de angariação conta com cinco fases essenciais:

1ª – Encontrar o imóvel disponível para venda direta através do cliente proprietário;

2ª – Entrar em contacto com o proprietário, oferecendo uma análise de mercado, de forma a precificar corretamente o imóvel, bem como o plano de marketing concretizado pela empresa, caso o cliente entregue a angariação à consultora;

3ª – Negociar valores e comissão adequados e justos para ambas as partes;

4ª – Organizar a documentação do imóvel, como caderneta predial, licença de utilização e certificado energético, sem o qual a empresa não pode promover o imóvel;

5ª – Realizar trabalho fotográfico e de vídeo, e por fim, publicar nos portais e redes sociais.

Com a situação pandémica, todo o processo de angariação sofreu alterações na EURO BROKERS. A angariação passou a ser maioritariamente do tipo reativo, recorrendo apenas aos portais imobiliários virtuais, e abandonando a prospeção externa. Nesta situação de mudança do mercado imobiliário, os clientes proprietários tornaram-se muito resistentes à venda dos imóveis devido à incerteza vivida em todos os setores, o que levou a consultora a procurar estratégias alternativas.

Neste sentido, e com o objetivo de fortalecer as redes sociais, a empresa investiu na promoção e atualização do *website*, bem como na compra de características adicionais no *software* de gestão X-IMO. Esta atualização permitiu o acesso a imóveis exclusivos publicados por proprietários.

Com o X-IMO e o *software* CASAFARI, que utiliza *data analytics* para a realização de modelos de avaliação automatizada, foi possível mitigar os efeitos nefastos da situação pandémica no volume de negócios da empresa. Foi nesta fase que se obteve um maior número de *clicks* e *leads* (visualizações), como se pode verificar na Figura 4, traduzindo sucesso na implementação das atualizações.

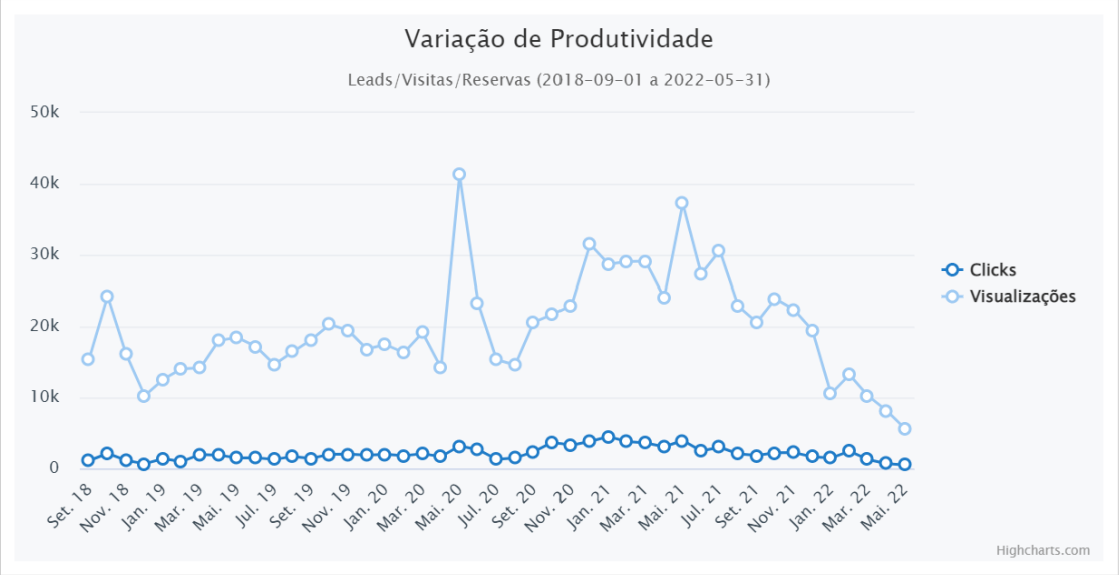


Figura 4 - Variação de *Clicks* e *Leads* desde 2018

Estima-se que o volume de negócios aumentou proporcionalmente às visualizações, como se verifica na Figura 5.

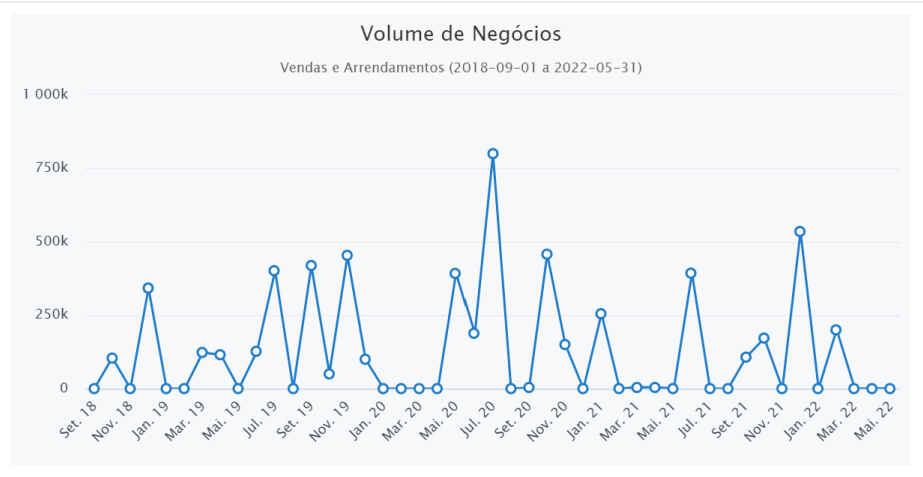


Figura 5 - Volume de Negócios de 2018 a 2022

Embora as estratégias de melhoria tenham sido aparentemente bem-sucedidas, a empresa procura, de uma forma constante, manter-se competitiva face às médias e grandes empresas de mediação imobiliária. É neste sentido e dimensão que o presente projeto se insere, uma vez que trará uma vantagem competitiva em relação à concorrência. Desta forma, a EURO BROKERS conseguirá oferecer ao cliente proprietário uma maior certeza numa variável que, normalmente, é de difícil previsão – o tempo de venda do imóvel.

3.2 Métricas da empresa

A EURO BROKERS, recentemente, passou por alterações no fluxo de trabalho, mudança na cultura organizacional e *re-design* de processos, que a obrigou a monitorizar alguns indicadores de desempenho com atenção.

Os principais indicadores utilizados pela empresa dizem respeito ao número de *leads*, visitas, reservas e negócios realizados. De salientar que a produtividade por colaborador (número de angariações) não é considerada como componente fundamental para a avaliação do mesmo.

No que diz respeito às métricas, é possível concluir, através da Figura 6, que é necessário um número elevado de *leads* para que estes se convertam em reservas e posteriores escrituras. São necessários cerca de 45 *leads* para se realizar uma reserva e duas reservas para se efetuar uma escritura. A taxa de conversão de *leads* para reservas é, então, de cerca de 2.2% e a de reservas para escrituras é de 50%.

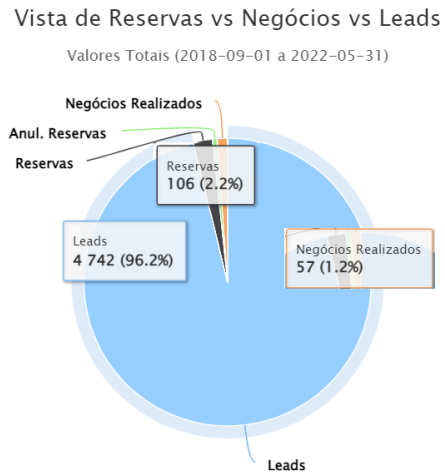


Figura 6 - Percentagens de *Leads*, Reservas e Negócios Realizados

O valor da taxa de conversão de 2.2% é bastante reduzido, no entanto pode ser explicado pela tecnologia e método que hoje está presente no ramo imobiliário. É cada vez mais fácil pedir informações sobre determinado imóvel, seja através dos portais virtuais, website da empresa ou mensagem, e por este motivo nem todos os contatos representam potenciais clientes.

No seguimento da explicação anterior, o mesmo acontece relativamente às visitas, como se pode visualizar na Figura 7, onde se verifica um número bastante menor de visitas, cerca de 1/3, comparativamente com o número de *leads*.

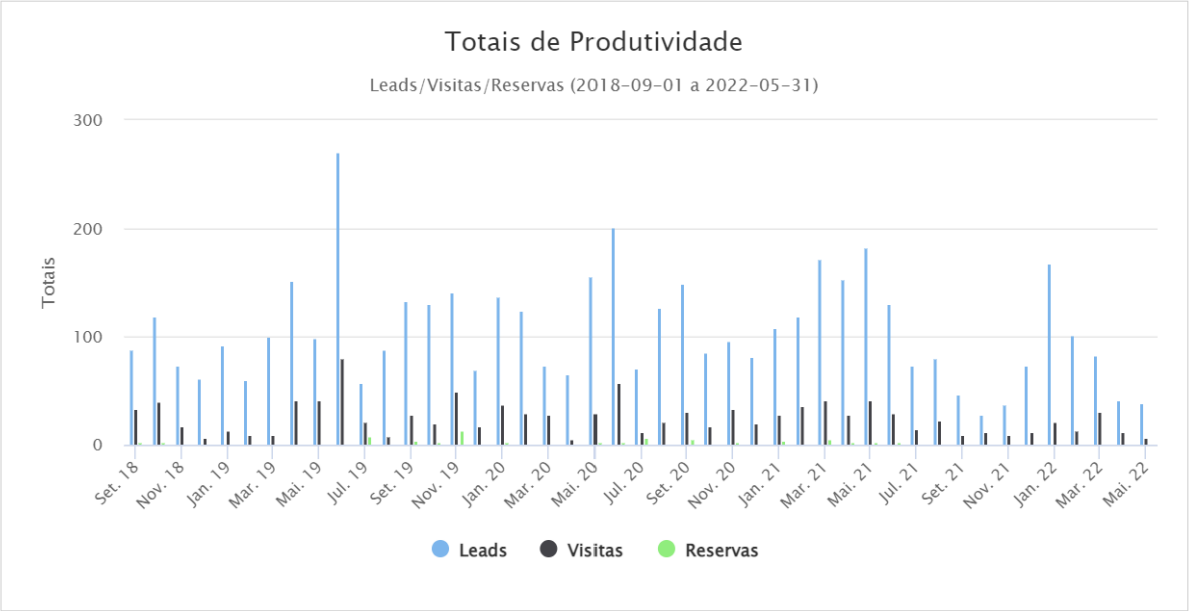


Figura 7 - Gráfico de Leads/Visitas/Reservas

Por outro lado, a percentagem de conversão de reservas para escrituras (50%) indica que foi feita uma boa correspondência entre a procura/necessidades dos clientes e a solução apresentada.

Numa perspetiva do processo de angariação, é possível perceber qual a tipologia que os clientes da EURO BROKERS mais procuram, como se verifica na Figura 8. Pode-se concluir que as tipologias mais procuradas são T2 (22,31%), T3 (21,02%) e T1 (18,35%).

Métrica de número de Leads por Tipologia (2018-09-01 a 2022-05-27)

Tipologia	Nº de Leads	% Leads
	215	4.53%
>=T10	3	0.06%
T0	280	5.9%
T0 DUPLEX	2	0.04%
T0+1	60	1.27%
T1	870	18.35%
T1 DUPLEX	6	0.13%
T1+1	150	3.16%
T1+2	6	0.13%
T2	1058	22.31%
T2 DUPLEX	66	1.39%
T2+1	135	2.85%
T2+1 DUPLEX	1	0.02%
T2+2	13	0.27%
T3	997	21.02%

Figura 8 - Número de *Leads* por Tipologia

3.3 Base de dados da empresa e respetivas variáveis

Como referido anteriormente, foram recolhidas 3 bases de dados estruturados, apenas relativas à empresa, (2 de clientes e 1 de imóveis) recorrendo ao *software* X-IMO, onde a EURO BROKERS realiza a gestão da empresa.

Estas bases de dados são consideradas neste estudo como *raw data*, pois é necessário um pré-processamento para preparar os dados para a modelação, sobretudo para a base de dados de imóveis. A análise das bases de dados nunca foi realizada anteriormente por parte da empresa, o que requereu, para a realização do presente projeto, um processo exaustivo e moroso, com o potencial de introduzir melhorias significativas na empresa.

3.3.1 Clientes Proprietários

A base de dados dos clientes proprietários conta com 300 observações e 10 variáveis: ID, nome, género, distrito, concelho, freguesia, localidade, data de inserção, número de imóveis e respetivas referências.

Para a realização da análise descritiva foi eliminada a coluna relativa ao nome dos proprietários, visto que os dados não irão ser agregados por esta variável. A percentagem de ocorrência é cerca de 60% para o sexo masculino e apenas 29% para o feminino. Os restantes 11% estão classificados como "outro". Em relação à localização dos proprietários foi apenas considerado o concelho, uma vez que a empresa trabalha maioritariamente no distrito do Porto, e, através desta análise, será possível extrair conclusões com resultados significativos.

Com o objetivo de realizar uma análise meramente descritiva, foram eliminadas observações sem qualquer tipo de conteúdo prático, como por exemplo, observações com o valor zero na variável "número de imóveis (por proprietário)", para que as mesmas não influenciassem os cálculos de média.

Através da análise visual à Figura 9, foi possível perceber que a base de dados continha alguns *outliers.* O cálculo dos quartis, máximo e mínimo do número de imóveis, permitiu identificar que o valor mínimo, a mediana e o primeiro quartil têm o valor de 1, enquanto que o terceiro quartil tem o valor de 2.

Uma vez que, para a análise dos *outliers*, é necessário calcular Q3 + 3 × (Q3-Q1), obteve-se um valor de 5. Desta forma, todos os valores superiores a 5 serão considerados *outliers*, e por esse motivo retirados da análise. A média de imóveis por proprietário é de 1,28.

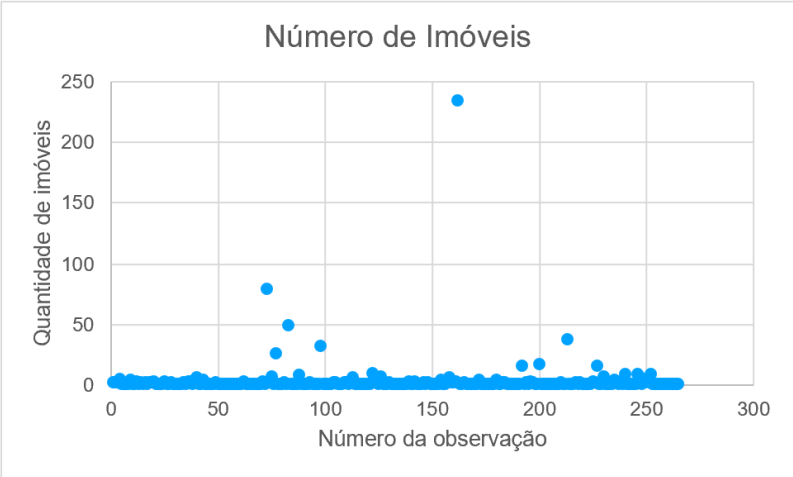


Figura 9 - Quantidade de imóveis por observação (por proprietário)

O concelho do Porto é o que apresenta o maior número de proprietários de imóveis. Importa referir que, nos casos em que o concelho não foi identificado, considerou-se a moda, ou seja, o concelho do Porto, dando origem aos números apresentados na Figura 10.

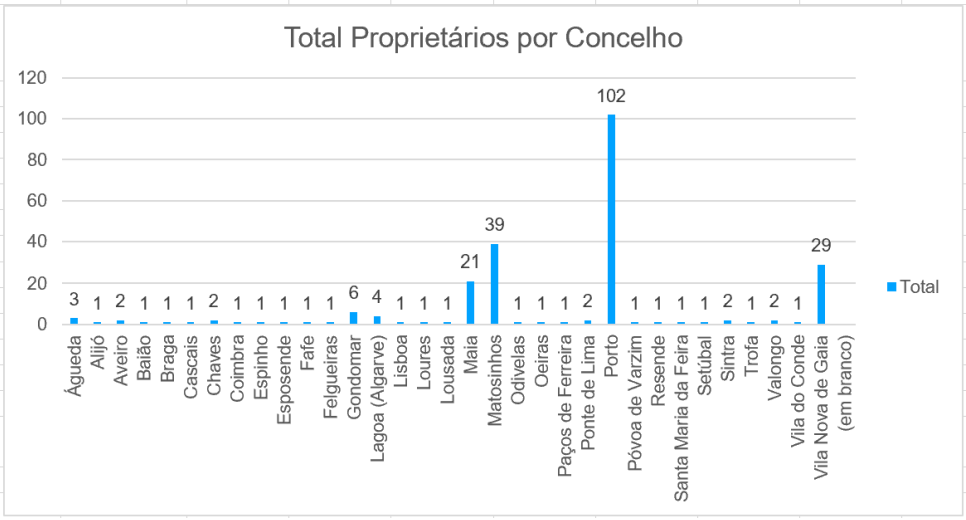


Figura 10 – Número de Proprietários por Concelho

De acordo com a Figura 11, verifica-se que o número de clientes proprietários da EURO BROKERS tem vindo a diminuir desde 2019. Embora esta métrica tenha vindo a diminuir, a empresa continua a corresponder às exigências dos clientes compradores, e a conseguir um volume de negócios que a permite manter-se em lucro.

Figura 11 - Número de Proprietário por Ano

Como mencionado anteriormente, reter um cliente é mais eficaz do que encontrar um novo, e sendo o portefólio da empresa preenchido, maioritariamente, por investidores, não tem sido necessário um processo de angariação muito exigente para manter a empresa com lucro.

Com o objetivo de compreender melhor a relação entre as variáveis presentes na base de dados proprietários, foi realizada uma análise recorrendo às *Association Rules*, explicadas anteriormente no Capítulo 2.

Utilizou-se um *support* mínimo de 0.5, para o algoritmo *Apriori*, e foram encontrados cinco conjuntos de itens frequentes e respetivos suportes, como se verifica na Figura 12.

```
    items                              support   count
[1] {distrito=Porto}                   0.8680851 204
[2] {n_imoveis=1}                      0.7829787 184
[3] {distrito=Porto, n_imoveis=1}      0.6978723 164
[4] {genero=M}                         0.5914894 139
[5] {genero=M, distrito=Porto}         0.5234043 123
```

Figura 12 - Conjuntos de Itens Frequentes e respetivos suportes da base de dados Proprietários

Uma vez que os conjuntos de itens frequentes deram posteriormente origem a regras com um valor de *support* baixo, os conjuntos [4] e [5] foram eliminados da análise, com o intuito de aprimorar o modelo.

Permanecendo apenas os conjuntos [1] a [3], e para a obtenção de regras mais adequadas, foi introduzido um novo parâmetro "minlen" igual a 2, para que fossem evitadas regras com apenas 1 item. Restringiu-se também as regras com um valor de confiança inferior a 0.6. As regras obtidas foram as da Figura 13.

```
    lhs                 rhs                 support   confidence coverage
[1] {n_imoveis=1}    => {distrito=Porto} 0.6978723 0.8913043  0.7829787
[2] {distrito=Porto} => {n_imoveis=1}    0.6978723 0.8039216  0.8680851
    lift     count
[1] 1.026748 164
[2] 1.026748 164
```

Figura 13 - Regras para os conjuntos de itens frequentes considerados

Conclui-se assim, que quando o número de imóveis é igual a 1, é provável que o proprietário seja do distrito do Porto. De salientar, que embora seja possível retirar esta conclusão da análise de associação, a métrica *lift* encontra-se muito próxima de 1, o que significa que as variáveis em questão são independentes, e por esse motivo, a regra considerada é pouco relevante para a análise geral.

3.3.2 Clientes Compradores

À semelhança do efetuado na base de dados dos clientes proprietários, também na dos clientes compradores foi realizada uma análise descritiva.

A base de dados dos clientes compradores conta com 4139 observações e 9 variáveis: ID, tipo, nome, género, data de inserção, situação, interesse, data da última ação, motivo da última ação. A variável nome foi, novamente, retirada pelo motivo referido anteriormente.

Ao contrário do que se verificou nos clientes proprietários, o género predominante nesta base de dados é o feminino, com uma percentagem de 50% face aos 42% do sexo masculino e os restantes 8% classificados como "outro".

Em relação à variável tipo, esta pode adquirir quatro denominações distintas: ativo, arquivo, inativo, potencial. Estas designações são atribuídas de acordo com o objetivo do cliente e o seu estado na empresa, ou seja, se tem visita marcada (ativo), se já realizou negócio (inativo), se pretende realizar mais (potencial) ou se já não necessita de qualquer serviço imobiliário (arquivo). A variável situação que descreve qual o estado do clientes, se tem visitas marcadas, se é necessário entrar em contato, entre outros, está intimamente ligada com a variável tipo, pelo que será descartada.

Como se verifica na Figura 14, 75% dos clientes compradores são potenciais clientes, o que significa que serão contactados quando existir um imóvel que corresponda às suas exigências. Nesta fase, é muito importante uma correta qualificação dos clientes, pois só assim é possível ir ao encontro das expetativas dos mesmos.

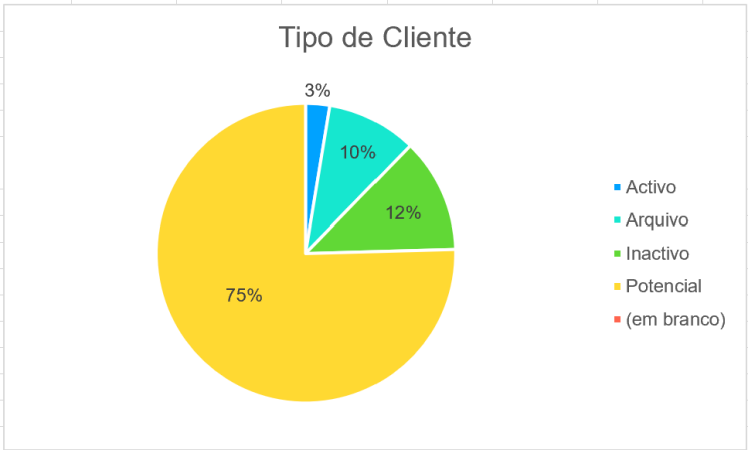


Figura 14 - Tipo de Cliente

A variável "Ano" demonstra um equilíbrio no número de clientes compradores ao longo dos anos, a partir do ano 2019. No primeiro ano da empresa o número foi menor, apenas 11% do total de clientes, no entanto tal estatística já era de esperar por ser uma empresa em começo de vida, com ainda poucos contactos.

Figura 15 - Número de Clientes por Ano

Relativamente à variável interesse, esta designa qual o objetivo dos clientes (comprar, arrendar, trespassar, entre outros), sendo esta maioritariamente compra e arrendamento, 60% e 32% respetivamente.

Por fim, as variáveis última ação e motivo da última ação serão eliminadas pois não contribuem de forma construtiva para a análise, uma vez que as ações podem ser de inúmeras formas, não havendo uma formalização das mesmas.

Através, mais uma vez, do método *Association Rules* aplicado à base de dados em questão, com um *support* mínimo de 0.5, foram encontrados dois conjuntos de itens frequentes: Interesse "Comprar" e Tipo "Potencial", onde o *support* adquire valores de 0.6145 e 0.7545, respetivamente. Para a obtenção de regras mais adequadas, também nesta análise foi introduzido o parâmetro "*minlen*" igual a 2.

Foram conseguidos os resultados da Figura 16:

```
    lhs                     rhs                   support   confidence
[1] {INTERESSE=Comprar} => {TIPO=Potencial}    0.5070082 0.8250098
[2] {TIPO=Potencial}    => {INTERESSE=Comprar} 0.5070082 0.6720051
    coverage  lift     count
[1] 0.6145481 1.093495 2098
[2] 0.7544708 1.093495 2098
```

Figura 16 – Regras obtidas no método *Association Rules* para a base de dados dos compradores

Embora o valor de confiança seja de 0.825 para a regra [1], o *lift* associado é muito próximo de um, o que significa, mais uma vez, que o antecedente e o consequente de regra são independentes.

Com o objetivo de analisar de uma forma mais profunda a relação entre as variáveis, foi diminuído o *support* mínimo para 0.3. Desta forma, foram identificados mais conjuntos de itens frequentes, e respetivos suportes, como se verifica na Figura 17.

```
    items                                 support   count
[1] {TIPO=Potencial}                      0.7544708 3122
[2] {INTERESSE=Comprar}                   0.6145481 2543
[3] {TIPO=Potencial, INTERESSE=Comprar}   0.5070082 2098
[4] {GÊNERO=F}                            0.4946834 2047
[5] {GÊNERO=M}                            0.4178347 1729
[6] {TIPO=Potencial, GÊNERO=F}            0.3755437 1554
[7] {INTERESSE=Arrendar}                  0.3250362 1345
[8] {TIPO=Potencial, GÊNERO=M}            0.3131948 1296
```

Figura 17 - Conjuntos de Itens Frequentes, e respetivos suportes após diminuição do *support*

Dos conjuntos de itens frequentes foram identificadas as regras presentes na Figura 18, cujo valor da métrica *lift* continua perto do valor 1.

```
    lhs                     rhs                   support   confidence
[1] {GÊNERO=M}          => {TIPO=Potencial}    0.3131948 0.7495662
[2] {GÊNERO=F}          => {TIPO=Potencial}    0.3755437 0.7591597
[3] {INTERESSE=Comprar} => {TIPO=Potencial}    0.5070082 0.8250098
[4] {TIPO=Potencial}    => {INTERESSE=Comprar} 0.5070082 0.6720051
    coverage  lift     count
[1] 0.4178347 0.9934994 1296
[2] 0.4946834 1.0062149 1554
[3] 0.6145481 1.0934948 2098
[4] 0.7544708 1.0934948 2098
```

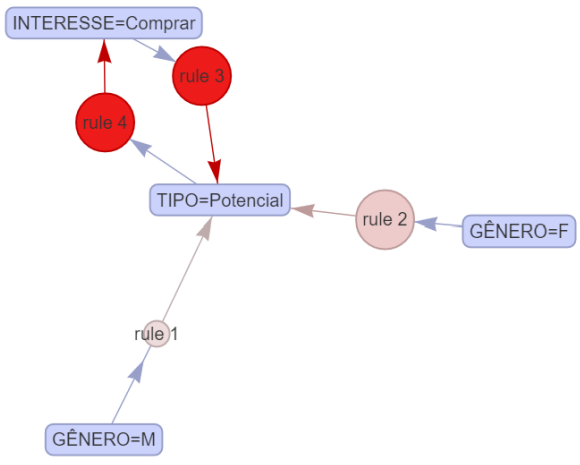


Figura 18 - Regras para os conjuntos de itens frequentes considerados, e respetivas métricas

Assim, e confirmando as conclusões retiradas na primeira análise, conclui-se que as variáveis presentes na base de dados são independente, embora seja possível perceber qual é a probabilidade de ocorrência entre elas. A titulo de exemplo, quando o Interesse é "Comprar", é muito provável que o Tipo associado seja "Potencial" (regra com o maior valor de confiança).

3.3.3 Imóveis

A base de dados correspondente aos imóveis representa o foco da análise, pois servirá de base para a modelação seguinte, com o objetivo de prever o tempo de venda de um imóvel. É, por este mesmo motivo, a mais importante e a que requer um maior trabalho a nível de pré-processamento dos dados.

Esta base de dados contém 1147 observações e 24 variáveis: ID, tipo, tipologia, referência, localização, área do terreno, área útil, área bruta, área bruta privativa, objetivo, estado, preço, data de angariação, exclusividade, classe energética, estado do negócio, comissão percentual, comissão fixa, transação da venda, data da transação, preço anterior, última alteração de preço, data retirado do disponível.

Os imóveis em carteira da EURO BROKERS são maioritariamente apartamentos (83%) localizados no distrito do Porto, sendo apenas 7% moradias. A empresa também possui terrenos, lojas e escritórios para venda, que serão retirados da análise, visto que o processo e tempo de venda não é comparável com o dos imóveis de habitação.

Para além da cidade do Porto (54%), e tal como se verifica na Figura 19, a empresa também tem uma forte presença em Lagos (20%) e Lisboa (9%).

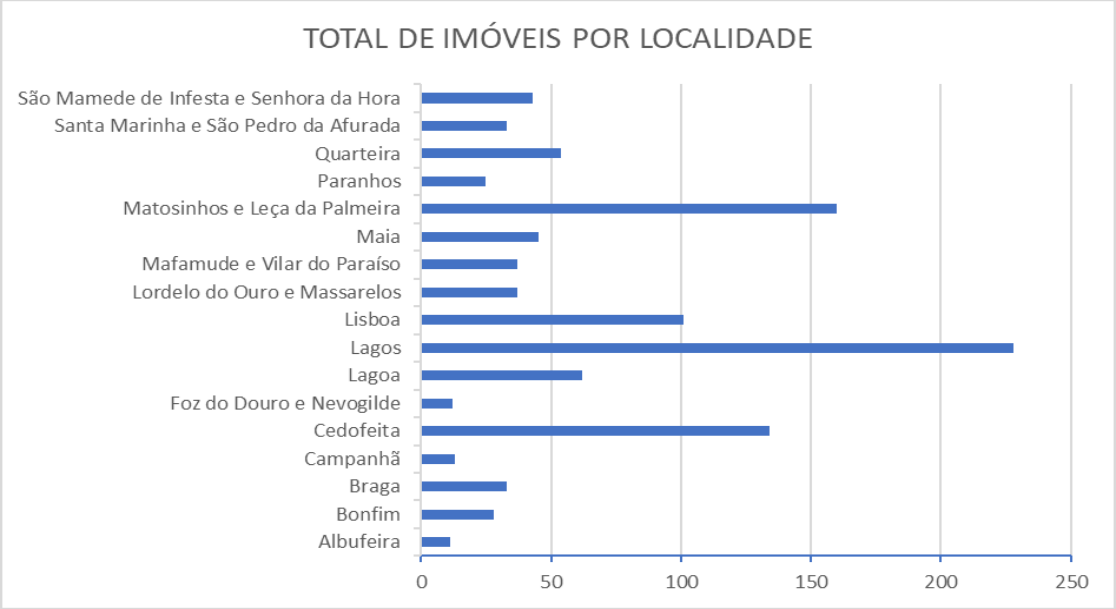


Figura 19 - Imóveis por Localidade

Tendo em conta que a variável área do terreno pode incluir jardim, arrumos, garagem, entre outros, e que cerca de 87% das observações se encontram em branco, esta foi desconsiderada no estudo do presente projeto.

A área útil diz respeito à soma de todas as áreas disponíveis para habitação, medidas a partir da parte interior das paredes. A área bruta é a área total do imóvel, medida pelo perímetro exterior das paredes e inclui as áreas da garagem, varandas, terraços e arrumos. Por fim, a

área bruta privativa de um imóvel refere-se a todas as áreas privadas/fechadas, sem contabilizar varandas abertas, garagens e arrumos. Esta métrica não é frequentemente utilizada no ramo imobiliário, pelo que será desconsiderada.

Como referido anteriormente, os tipos de imóveis maioritariamente procurados pelos clientes da EURO BROKERS, são os T1, T2, e T3, pelo que estes representam a maior parte da angariação efetuada pela empresa. Paralelamente, os T0 são procurados para investimento, não sendo contabilizados na análise de *leads* comuns.

Os *missing values* (4 observações) desta variável serão substituídos pela moda dos imóveis da respetiva localidade.

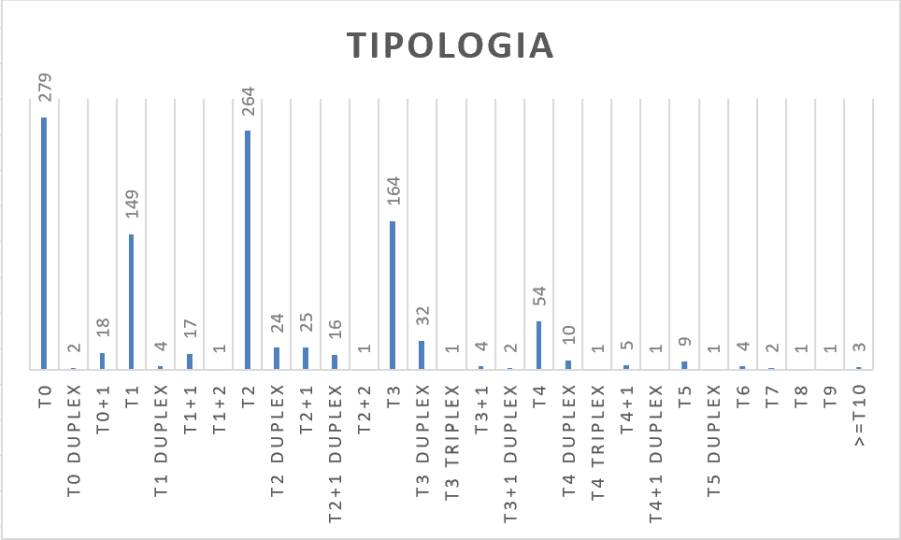


Figura 20 - Total de Imóveis por Tipologia

Tendo em conta que o objetivo principal deste projeto se prende com a previsão do tempo de venda de um imóvel, a variável objetivo (venda, trespasse ou arrendamento) não será tida em consideração, assim como todas as observações relativas a arrendamentos ou trespasses.

Como se pode verificar na Figura 21, os imóveis remodelados representam 38% dos imóveis em carteira, visto que, de um modo geral, a compra e remodelação de um imóvel usado trata-se de um melhor investimento do que a compra de um imóvel novo. Os restantes estados, com a exceção dos imóveis para remodelar (que são rapidamente convertidos no estado remodelado), apresentam valores semelhantes entre si.

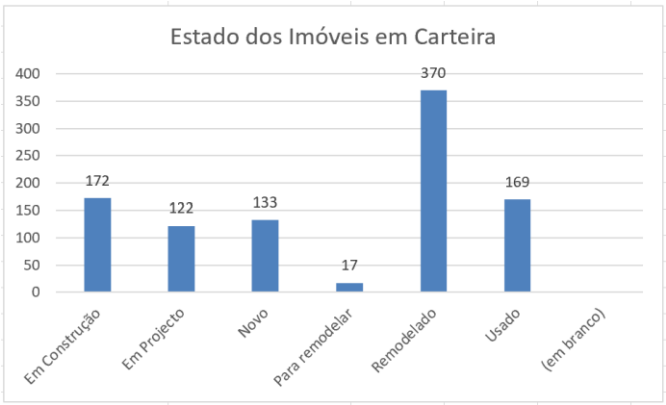
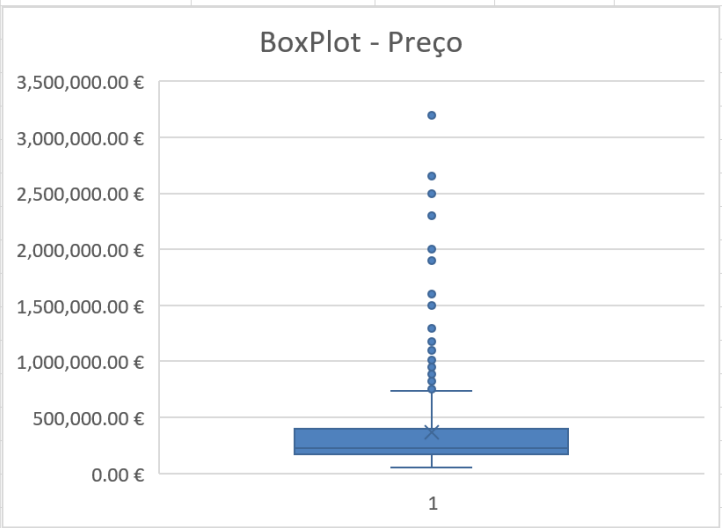


Figura 21 - Quantidade de Imóveis por Estado de Utilização

A aquisição de um imóvel novo não permite uma valorização adicional, estando apenas dependente da valorização da zona onde se insere, enquanto que um imóvel usado permite ao

proprietário valorizá-lo, através da sua remodelação, que leva a um aumento significativo do seu valor independentemente da zona.

Uma das variáveis mais relevantes para a análise é o preço do imóvel, pois está, numa perspetiva geral, muito relacionado com o tempo de venda do mesmo. Através da análise descritiva desta variável, recorrendo a um *boxplot*, foi possível perceber que a base de dados contém alguns *outliers* (de preço superior a 1.082.100€) que devem ser eliminados para o cálculo da média.



Medida	Valor em €
Mínimo	55,000.00 €
Q1	169,000.00 €
Mediana	229,000.00 €
Q3	397,275.00 €
Máximo	3,200,000.00 €
Q3 + 3* (Q3-Q1)	1,082,100.00 €
Média c/ *outliers*	367,128.86 €
Média s/ *outliers*	304,627.41 €

Figura 22 - *BoxPlot* e medidas relativas ao Preço

O tempo de venda (em meses) poderá ser estimado a partir das variáveis data de angariação e data de retirado de disponível (subtraindo a primeira da segunda). Desta forma, descartaram-se, para este estudo, as duas variáveis utilizadas para o cálculo.

Analisando a nova variável, é possível concluir que a mesma não apresenta *outliers* (Figura 41 do Anexo A) e a média de meses que uma casa demora a ser vendida é de aproximadamente 11. No entanto, o tempo de venda é influenciado por vários fatores, que serão analisados posteriormente, e que explicarão este valor tão elevado.

Os imóveis já vendidos (e em que o tempo de venda ≠ 0) correspondem a 93% da base de dados, sendo os restantes (aproximadamente 70) desconsiderados.

Relativamente à variável exclusividade, esta não adiciona nenhuma informação extra à análise, pois 99% dos imóveis encontram-se em regime de não exclusividade, ou seja, para além da EURO BROKERS, outras empresas de mediação imobiliária também promoveram o imóvel. Tendo em consideração esta informação, a variável exclusividade foi desconsiderada.

A crescente valorização de imóveis com melhor certificação energética torna esta variável importante na determinação do tempo de venda de um imóvel. Neste sentido, e visto que há um elevado número de observações sem certificação energética, foi utilizada, numa primeira fase, a moda de cada subgrupo da variável estado (Figura 18). Posteriormente, de forma a obter melhores resultados, foi utilizado o método *k-NN*, explicado no próximo capítulo.

	Estado do Imóvel					
Classe Energética	Em Construção	Em Projeto	Novo	Para remodelar	Remodelado	Usado
A	7	0	27	0	3	1
A+	2	0	0	0	0	2
B	0	0	3	0	4	9
B-	1	0	23	0	9	11
C	0	0	0	1	43	65
D	0	0	0	2	34	33
E	0	0	0	3	21	12
F	0	0	0	0	2	4
G	1	0	0	0	2	5
Moda	**A**	**-**	**A**	**E**	**C**	**C**

Figura 23 - Classe Energética em função do Estado do Imóvel

Assim, e tendo em conta que os imóveis em projeto ainda não dispõem de classificação energética, 39% dos imóveis descritos na base de dados são de classe C e 28% de classe A, como se pode inferir a partir dos dados expostos na Figura 24.

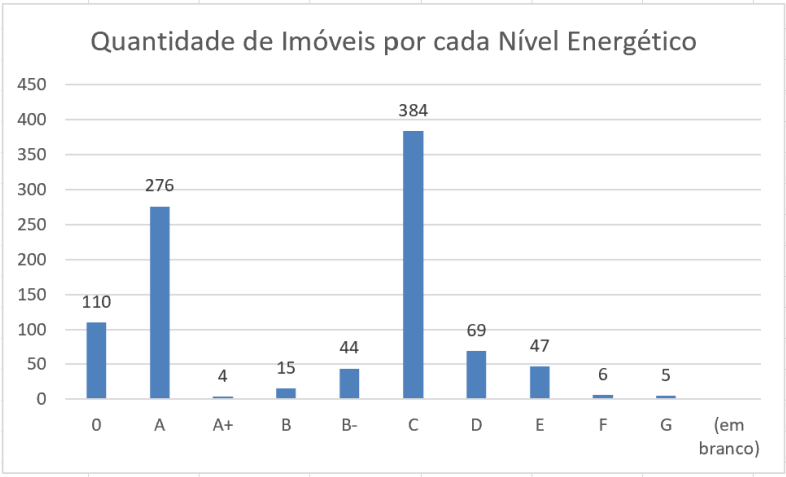


Figura 24 - Quantidade de Imóveis por cada Classe Energética

A variável estado de negócio refere-se aos imóveis disponíveis para venda e aos indisponíveis (cancelado/retirado de venda ou indisponível/negócio realizado), onde os disponíveis representam 91% dos dados. Assim, e visto ambas as variáveis terem o mesmo contributo, o estado de negócio foi removido.

Relativamente à comissão, a EURO BROKERS aplica, geralmente, uma comissão percentual (5% + IVA) e, em casos raros, uma comissão fixa. Para este estudo, nos casos em que foi aplicada uma comissão fixa, esta foi convertida no seu valor percentual.

Tendo em conta que não existia informação suficiente relativa às últimas quatro variáveis (transação da venda, data da transação, preço anterior e última alteração de preço), estas foram desconsideradas na análise.

Finalmente, foi realizada uma análise recorrendo a *Association Rules*, à semelhança do realizado anteriormente para as demais bases de dados. No entanto, foi necessário recorrer à discretização das variáveis área útil, preço, comissão percentual e número de meses. Foi, para isso, utilizado o método *Equal-width*, com um número de *bins* calculado através da raiz quadrada da dimensão da base de dados – 30 *bins*.

Para um valor mínimo de suporte de 0.4, foram encontrados 3 conjuntos de itens frequentes, correspondentes à variável Apartamento e Classe Energética C, separadamente, e em conjunto. Posteriormente, foram encontradas as regras da Figura 25.

```
    lhs                      rhs                    support   confidence
[1] {classe_energetica=C} => {tipo=Apartamento}    0.4713303 0.9903614
[2] {tipo=Apartamento}    => {classe_energetica=C} 0.4713303 0.5229008
    coverage  lift     count
[1] 0.4759174 1.098722 411
[2] 0.9013761 1.098722 411
```

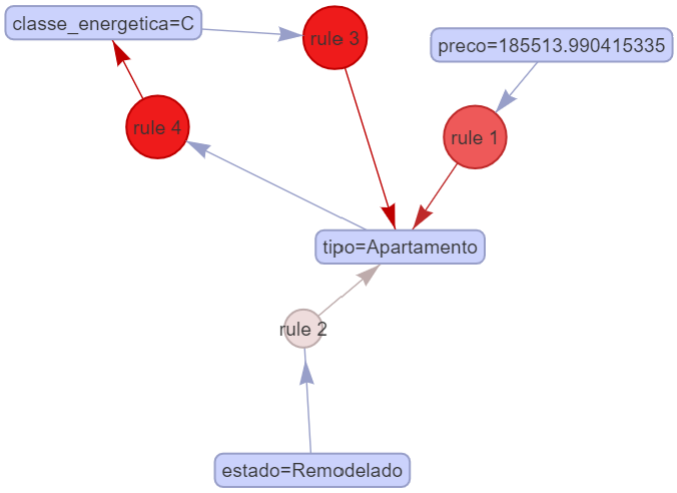


Figura 25 - Regras obtidas do método *Association Rules*, e respetivas métricas, para a base de dados Imóveis

A regra [1] tem uma confiança de aproximadamente 1, o que significa que quando o imóvel é de classe C, é muito provavel que seja um apartamento. Mais uma vez, existe independência entre o antecedente e o consequente da regra, mostrando que esta é pouco relevante.

Testando o modelo para um número mínimo de suporte mais baixo, foi possível perceber que praticamente todas as regras identificadas tinham o valor de *lift* próximo de 1, significando, uma vez mais, uma baixa correlação entre variáveis, e uma elevada independência.

4 Modelo e Soluções

Como forma de desenvolver soluções para os objetivos propostos, foram comparados diferentes modelos de previsão (árvores de regressão, redes neuronais, *nearest neighbor regression* e *random forest*), recorrendo ao *software* R e ao *Microsoft Excel* para a realização de todas as análises e elaboração dos algoritmos.

Foi efetuado um processo elaborado de pré-processamento, como forma de melhorar a qualidade do modelo desenvolvido. O objetivo principal do projeto é a previsão do tempo de venda de um imóvel (em meses), recorrendo para isso a uma série de variáveis independentes (tipologia, localização, entre outros). A correta previsão da variável permitirá o aumento da credibilidade da empresa no momento da angariação de imóveis.

Métricas como o erro quadrado médio, erro raiz quadrada média e erro médio absoluto ditarão qual o modelo a implementar no futuro, e consequentemente, se o objetivo foi cumprido.

4.1 Preparação dos dados

Para uma modelação correta e um aumento da qualidade das soluções desenvolvidas, torna-se necessário preparar os dados para a mesma. Esta preparação inclui um conjunto de técnicas para transformar os dados brutos em formatos de fácil compreensão e analisáveis. Existem cinco passos envolvidos neste processo:

- Limpeza de dados – envolve a correção de dados vazios, *outliers* e inconsistências (parte deste processo foi realizado anteriormente, e descrito no capítulo 3);
- Integração de dados – consiste na combinação de várias bases de dados, que neste projeto, não foi executado;
- Transformação de dados – envolve a generalização dos dados, nomeadamente normalização dos mesmos;
- Redução dos dados – consiste na redução da base de dados quando esta é demasiado numerosa. Não foi necessário executar este passo no presente projeto;
- Discretização dos dados – consiste na conversão de variáveis contínuas em variáveis discretas, criando um número limitado de estados possíveis (utilizada em variáveis numéricas).

No primeiro passo (limpeza de dados), foi realizada uma verificação de duplicados na base de dados. Inicialmente, a base de dados aparentava conter cerca de 90 observações repetidas (imóveis com as mesmas características). No entanto, estes diferiam na referência do imóvel, tratando-se de empreendimentos com vários imóveis. Assim, nenhuma observação foi eliminada neste passo.

Após o processamento realizado no capítulo 3, eliminando variáveis sem contribuição analítica e observações irrelevantes, os valores iguais a zero (nas variáveis de áreas e

comissão) foram alterados para *NA*, permitindo ao *software* R interpretar esta informação e utilizar os dados nos passos subsequentes.

De forma a tratar os *missing values* presentes nas variáveis comissão, classe energética, área útil e área bruta, foi utilizado o método *k-NN*, que utiliza a semelhança de observações para prever os valores de novos dados. Isto significa que o novo valor é obtido através do cálculo da média das observações mais próximas (Vijay Kotu 2014). Embora represente uma grande simplicidade no seu método, apresenta uma elevada eficácia.

Em primeiro lugar, é necessário realizar o cálculo da distância entre o novo ponto e os pontos semelhantes, recorrendo ao método *Manhattan*, através da equação seguinte:

$$\sum_{i=1}^{k} |x_i - y_i| \tag{4.1}$$

Onde:

x_i, é o valor da observação i
y_i, é o valor da nova observação i
k, é o número de vizinhos observados

Na equação (4.1), efetua-se o somatório das diferenças absolutas entre o novo ponto e os semelhantes, até ao valor k – número de vizinhos que são observados até atribuição de valor à nova observação. De salientar que foi utilizada a função *k-NN* no *software* R, que integra este cálculo, e foi definido um k = 10.

Para verificação do parâmetro k, foi realizado um teste de *accuracy* (Figura 26), concluindo-se que k = 6 confere uma maior exatidão na previsão dos valores das variáveis mencionadas anteriormente – comissão, classe energética e áreas.

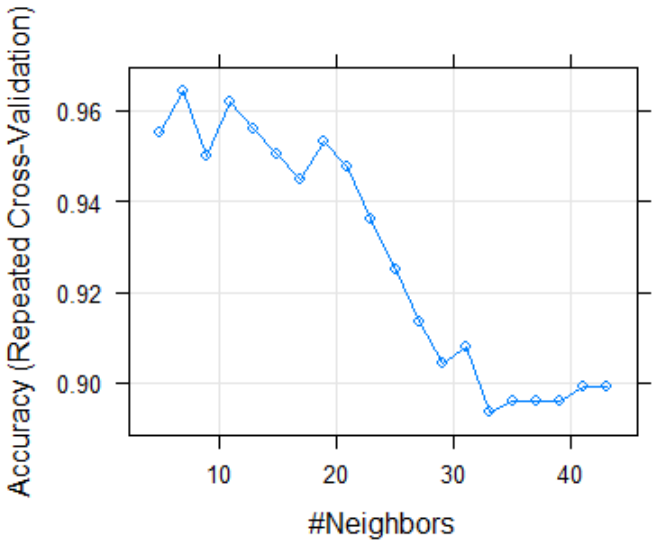
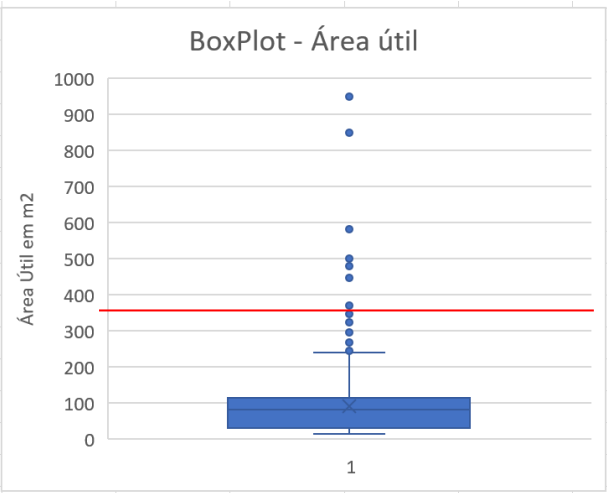
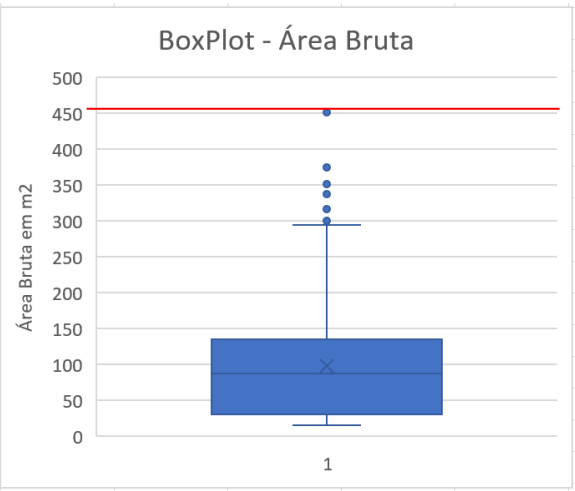


Figura 26 - *Accuracy* em função do número de vizinhos

A área bruta não pode ser inferior à área útil; assim, no preenchimento dos *missing values* destas duas variáveis, torna-se necessário ter este parâmetro adicional em conta. Após a previsão de todos os valores, foi possível contar 24 observações (2,5% das observações) em que tal acontecia, tendo estas sido eliminadas.

Estando todas as variáveis preenchidas, foi necessário analisar a área útil e bruta com o objetivo de encontrar *outliers*, recorrendo, novamente, aos *boxplots* e quartis. Foram encontrados poucos *outliers* (cerca de 7 – representados pelos pontos a cima da linha vermelha) na variável área útil que foram eliminados, e nenhum na área bruta, como se verifica na Figura 27.





Medida	Área Útil
Mínimo	15.00
Q1	30.00
Mediana	81.22
Q3	114.00
Máximo	949.00
Q3 + 3* (Q3-Q1)	366.00
Média c/ *outliers*	90.58
Média s/ outliers	84.92

Medida	Área Bruta
Mínimo	15.00
Q1	30.00
Mediana	87.30
Q3	135.65
Máximo	451.60
Q3 + 3* (Q3-Q1)	452.60
Média c/ *outliers*	97.85

Figura 27 - *BoxPlots* e respetivas medidas da Área Útil e Bruta

O passo seguinte consiste na normalização das variáveis numéricas – área útil, área bruta, preço, comissão percentual e número de meses. Esta etapa converterá variáveis com valores compreendidos na ordem das centenas (área) ou milhares/milhões (preço) em valores na mesma ordem.

Foi utilizado o método *Z-Score*, onde é feita a normalização de cada valor, de tal forma que a média de todos os valores seja igual a 0 e o desvio padrão igual a 1. Assim, para o cálculo do novo valor normalizado, é realizada uma subtração entre o valor original e a média do grupo, que é posteriormente dividido pelo desvio padrão também do grupo.

De forma a obter a correlação entre variáveis, a base de dados foi dividida em variáveis numéricas e variáveis categóricas. Nas variáveis numéricas foi calculado o coeficiente de correlação de *Pearson*, de acordo com a equação seguinte, que varia entre -1 e 1.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}} \tag{4.2}$$

Onde:

x_i, é o valor da observação i da variável x
y_i, é o valor da observação i da variável y
n, é o número total se observações

$\bar{x}$, é a média aritmética da variável x
$\bar{y}$, é a média aritmética da variável y
ρ, é o coeficiente de *Pearson*

Em relação aos valores do coeficiente, estes podem ser analisados da seguinte forma, independentemente do sinal dos mesmos (positivo ou negativo) (Benesty et al. 2009).

- 0,9 ou superior indica uma correlação muito forte;
- 0,7 a 0,9 indica uma correlação forte;
- 0,5 a 0,7 indica uma correlação moderada;
- 0,3 a 0,5 indica uma correlação fraca;
- 0 a 0,3 indica uma correlação negligenciável.

Analisando a matriz relativa às variáveis numéricas (Figura 28), comprova-se que a área útil e a área bruta têm uma correlação muito forte, e por esse motivo uma delas pode ser eliminada da análise. Tendo em conta que a métrica mais utilizada no ramo diz respeito à área útil, esta será mantida.

A maioria das restantes correlações são negligenciáveis ou fracas. As exceções dizem respeito às correlações entre área útil/preço, área útil/comissão e área bruta/comissão (correlações moderadas).

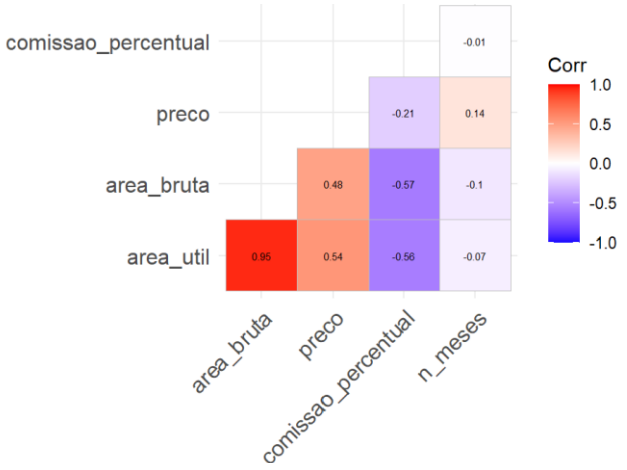


Figura 28 - Matriz de Correlação de *Pearson*

A partir da mesma análise, é possível verificar que nenhuma das variáveis de entrada tem uma correlação forte com a variável de saída, demonstrando a complexidade da previsão do tempo de venda do imóvel em meses. Desta forma, outras variáveis terão de ser tidas em consideração em trabalhos futuros.

Relativamente às variáveis categóricas, foi utilizado o *Cramér's V* que mede a correlação entre duas variáveis categóricas, de acordo com as seguintes equações, e varia entre 0 e 1, onde 0 significa nenhuma correlação e 1 uma correlação perfeita (Benesty et al. 2009).

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{4.3}$$

Onde:

ϕ_c, é o coeficiente de *Cramér's V*
χ^2, é o valor do *Pearson chi-square*
N, é o tamanho da amostra
k, é o menor número de categorias de qualquer variável

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.4}$$

Onde:

χ^2, é o valor do *Pearson chi-square*
n, é o tamanho da amostra
O_i, é o valor original da observação i
E_i, é o valor esperado da observação i

Quanto maior a diferença entre O_i (observações originais) e E_i (observações esperadas), maior será o qui-quadrado. Os cálculos foram realizados recorrendo ao *software* R, que atribui significância a cada valor, através da comparação com o valor crítico de 0,05.

A análise da matriz de correlação de *Cramér's V* (Figura 29) demonstra a existência de uma forte correlação entre o estado dos imóveis e a sua localização, sendo este o único coeficiente entre duas variáveis superior a 0,6 (Bergsma 2013).

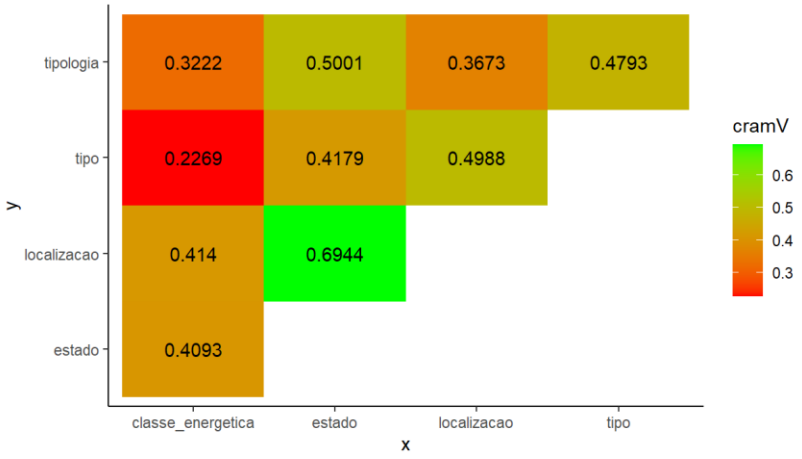


Figura 29 - Matriz de correlação de *Cramer's V*

De acordo com esta análise, e tendo em conta a correlação forte entre estado/localização e a correlação moderada entre estado/classe energética, a utilização da variável estado no preenchimento dos *missing values* da classe energética é a mais adequada.

A etapa de redução de dados não foi considerada neste projeto, pois a base de dados não é muito extensa, e se fosse realizado um *sampling* poderia perder-se informação crucial para a execução dos modelos de previsão.

4.2 Modelos de Previsão

A maioria dos modelos de previsão não consegue prever a variável de saída com sucesso, quando existem variáveis categóricas na sua composição. Por esse motivo, antes de iniciar os modelos, foi necessário transformar essas variáveis em *dummy variables*.

As *dummy variables* são variáveis fictícias que adquirem o valor 1 ou 0, consoante a sua presença ou não na observação. A título de exemplo, a variável estado é decomposta em seis variáveis fictícias, que correspondem aos valores que esta pode adquirir ao longo das observações: construção, projeto, novo, para remodelar, remodelado e usado.

Após esta transformação, o número de variáveis aumentou para 84 no total, representadas na Figura 46 do Anexo A, e as variáveis originais foram eliminadas, e os nomes das variáveis fictícias alterados para serem de melhor compreensão.

Posteriormente, foi realizada uma nova análise de correlação, onde se verificou que algumas das "novas" variáveis poderiam ser consideradas colineares pelo seu elevado coeficiente de correlação, o que significa que o comportamento de uma é completamente explicado pela outra. A variável prédio estava explicada pela variável T6, a variável projeto pela variável Matosinhos e as variáveis T0 e Lagoa pela variável comissão. Assim, estas foram eliminadas da análise.

Ainda nesta fase, e com o objetivo de reduzir o número de *dummy variables* em análise, foram retiradas aquelas cuja contribuição seria insignificante. O critério utilizado foi eliminar as variáveis que tivessem menos de 5 observações na base de dados.

Posteriormente, foi necessário perceber se existia uma correlação linear entre o número de meses e todas as variáveis quantitativas independentes originais – comissão percentual, preço

e área útil. Para isso, recorreu-se a gráficos de dispersão (Figura 30), onde foi traçada uma linha de tendência acompanhada com o respetivo R^2.

O R^2 é das métricas mais importantes no que diz respeito à linearidade entre variáveis, pois representa a proporção da variância da variável dependente que pode ser explicada pelas independentes. Um valor próximo de 1 significa que é explicada praticamente toda a variância da variável independente.

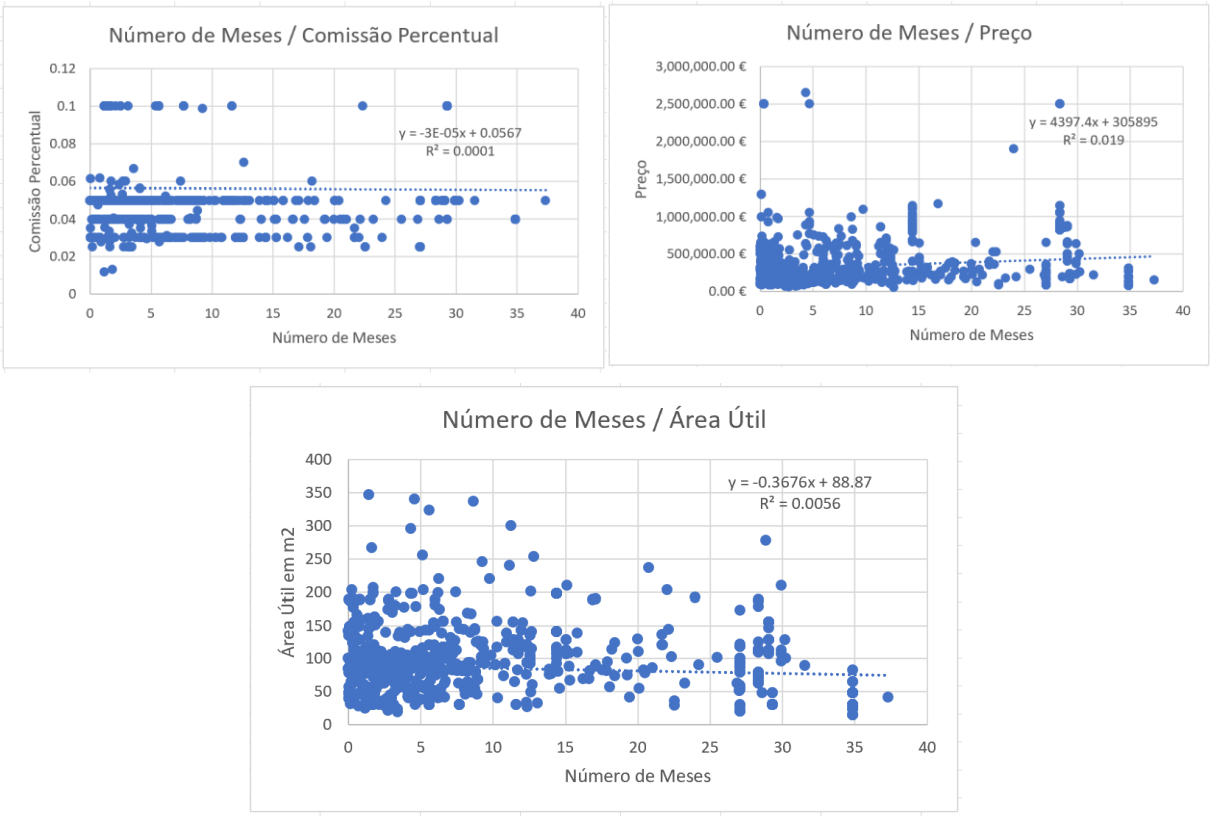


Figura 30 - Gráficos de Dispersão Comissão, Preço e Área Útil

Assim, foi possível constatar que a relação linear existente entre variáveis é extremamente fraca, o que comprova uma vez mais, que as variáveis presentes na base de dados não explicam por completo a variável a prever: tempo de venda.

Por último, a base de dados foi dividida em *training sample* e *testing sample*, para a validação do modelo de previsão, permitindo simular o desempenho do mesmo em dados ainda não fornecidos ao modelo (*testing sample*). Esta divisão é realizada na proporção 75/25 (*training sample/testing sample*) (Vijay Kotu 2014).

4.2.1 Árvores de Regressão

As árvores de regressão são algoritmos que utilizam decisões como recursos para representar o resultado previsto numa estrutura semelhante a uma árvore. Esta é gerada quando, em cada nó de decisão, é realizado um teste ao valor de alguma variável de entrada. Os nós terminais contêm os valores da variável de saída.

Este método é uma variante das árvores de classificação, projetado para aproximar funções de valor real, em vez de serem utilizados métodos de classificação.

Inicialmente, todas as observações do *training sample* são agrupadas numa só partição, e posteriormente, o algoritmo aloca os dados em duas ramificações, utilizando todas as divisões binárias possíveis em cada campo. A divisão selecionada é aquela que minimiza a soma dos

desvios quadrados da média nas duas partições separadas. Esta regra é aplicada a cada nova ramificação. Este processo finda quando cada nó atinge o tamanho mínimo imposto, tornando-se assim, um nó terminal (Loh 2011).

O primeiro passo é a construção da árvore de regressão tanto maior quanto possível, tendo em consideração um possível *overfitting*, uma vez que a dimensão da base de dados não é muito vasta. Para reduzir o impacto deste evento, é necessária a utilização do método *Cross Validation* (explicado detalhadamente mais à frente).

Utiliza-se assim, para a elaboração da árvore, um parâmetro de complexidade (*CP*) reduzido, neste modelo em específico de 0,001. Isto significa, de um modo geral, que serão realizadas ramificações na árvore desde que o R^2 geral do modelo aumente pelo menos no valor especificado do *CP*.

Os resultados obtidos encontram-se na Figura 31. Para cada valor do *CP* existe um erro associado, e por isso é necessário escolher o *CP* com o qual se obtém um erro mais baixo. Este processo é geralmente denominado por "poda da árvore", onde é encontrado o valor ótimo para o *CP*, ao qual está associado o menor erro "*xerror*" – que representa o erro nas observações dos dados de validação cruzada.

```
Variables actually used in tree construction:
 [1] A                  area_util          braga
 [4] C                  cedofeita          comissao_percentual
 [7] construcao         D                  lagos
[10] matosinhos         novo               preco
[13] t_tres             t_um               usado

Root node error: 922/923 = 0.99892

n= 923

           CP nsplit rel error  xerror     xstd
1  0.13392631      0   1.00000 1.00307 0.037195
2  0.07508935      1   0.86607 0.86948 0.034830
3  0.03867846      2   0.79098 0.81731 0.037344
4  0.03430062      3   0.75231 0.79869 0.037621
5  0.02969818      4   0.71801 0.77812 0.037275
6  0.02382593      5   0.68831 0.75250 0.036554
7  0.01560780      6   0.66448 0.71399 0.034454
8  0.01181016      9   0.61766 0.67830 0.034391
9  0.01119600     10   0.60585 0.67807 0.034870
10 0.00957514     11   0.59465 0.67508 0.034731
11 0.00873548     12   0.58508 0.65661 0.034561
12 0.00848635     13   0.57634 0.65516 0.034825
13 0.00590351     14   0.56785 0.61872 0.033731
14 0.00541033     15   0.56195 0.63294 0.034725
15 0.00531629     16   0.55654 0.62622 0.034826
16 0.00522893     18   0.54591 0.62431 0.034689
17 0.00494752     20   0.53545 0.62801 0.034865
```

Figura 31 - Possíveis CP's para Árvore de Regressão

Assim, e recorrendo ao *software* R, nomeadamente às funções "*rpart*" e "*prune*", a árvore de regressão foi construída e podada, sendo possível observar nos nós terminais a quantidade de imóveis e respetivo tempo de venda de acordo com determinadas condições.

De acordo com os resultados obtidos, representados na Figura 42 do Anexo A, pode concluir-se que:

- Existem 106 imóveis não usados, com um preço superior a 179.500€, comissão superior a 2,9% e área compreendida entre 85,75 m^2 e 143,95 m^2, cujo tempo de venda é de 9,96 meses. Existem apenas 8 imóveis do tipo usado, com tempo de venda de aproximadamente metade, 4,68 meses.
- O preço é um dos fatores que mais influencia o tempo de venda. Existem 72 imóveis cujo preço é inferior a 234.000€, tendo sido vendidos em 14,39 meses, aumentando para 21,66 meses quando o preço do imóvel é superior ao mencionado (26 imóveis analisados). Esta análise teve como condições, os imóveis não estarem em construção,

não serem usados, não serem em Cedofeita, terem área inferior a 116,82 m^2, não serem T1 nem T3 e o valor da comissão ser superior a 4,5%.

- Em Braga, o tempo de venda é significativamente superior, quando comparado com as restantes localizações, cerca de 34 meses. No entanto, é de referir que os imóveis considerados se encontram em fase de construção, o que atrasa o processo de venda.
- Existem 83 imóveis de preço inferior a 179.500€, percentagem de comissão superior a 2,6%, de classe energética D, com tempo de venda de 2,86 meses. Desta forma, é possível concluir que a classe energética influencia o tempo de venda do imóvel, uma vez que, mantendo todas as condições exceto a classe energética, o imóvel demora aproximadamente 6 meses a ser vendido.
- Relativamente à área útil, existem 9 imóveis com valores compreendidos entre 143,95 e 193,30 m^2, com um tempo de venda de 10,46 meses; para áreas inferiores a esse valor, existem 38 imóveis e o tempo de venda diminui para 3,24 meses. As condições presentes são as mesmas do ponto anterior, à exceção da classe energética.

A previsão foi realizada recorrendo à função "*predict*" e foi utilizada a *testing sample*, sendo os valores obtidos pela previsão comparados com os valores originais.

Com o objetivo de testar a capacidade do modelo, recorreu-se aos erros associados à previsão de problemas de regressão como demonstrado na Tabela 2.

Tabela 2 - Métricas de Avaliação das Árvores de Decisão

	Erro Quadrado Médio	Erro Raiz Quadrada Médio	Erro Médio Absoluto
Árvores de Decisão	56.14	7.49	5.01

Analisando estes valores, conclui-se que não são satisfatórios, pois um erro raiz quadrada médio de 7,49 meses significa que o modelo prevê cerca de 7,5 meses a mais ou a menos comparativamente com o valor real. Posto isto, e analisando com mais profundidade a base de dados após pré-processamento, foi possível constatar que os imóveis localizados fora do Porto têm um tempo de venda muito superior aos localizados nesta cidade. Assim, e com o objetivo de melhorar o modelo de previsão, foi utilizada uma sub-base de dados, composta apenas por imóveis na cidade do Porto, excluindo assim Lisboa, Braga, Aveiro e Algarve.

Na repetição do modelo de previsão foi possível obter a árvore de regressão presente na Figura 32, onde é possível retirar as seguintes conclusões:

- Para um preço inferior a 179.750€, um imóvel usado tem um tempo de venda mais baixo do que um imóvel não usado (2,8 e 5 meses, respetivamente).
- Para imóveis com o preço superior a 179.750€, em construção e de tipologia T2, o tempo de venda é de 7,09 meses.
- Para as mesmas condições iniciais: preço superior a 179.750€, sem estar em construção e com área útil inferior a 86,6 m^2, um imóvel com classe energética C é vendido mais rapidamente do que um que não tenha esta classe energética, quase quadruplicando o tempo de venda – de 1,82 para 6,95 meses.

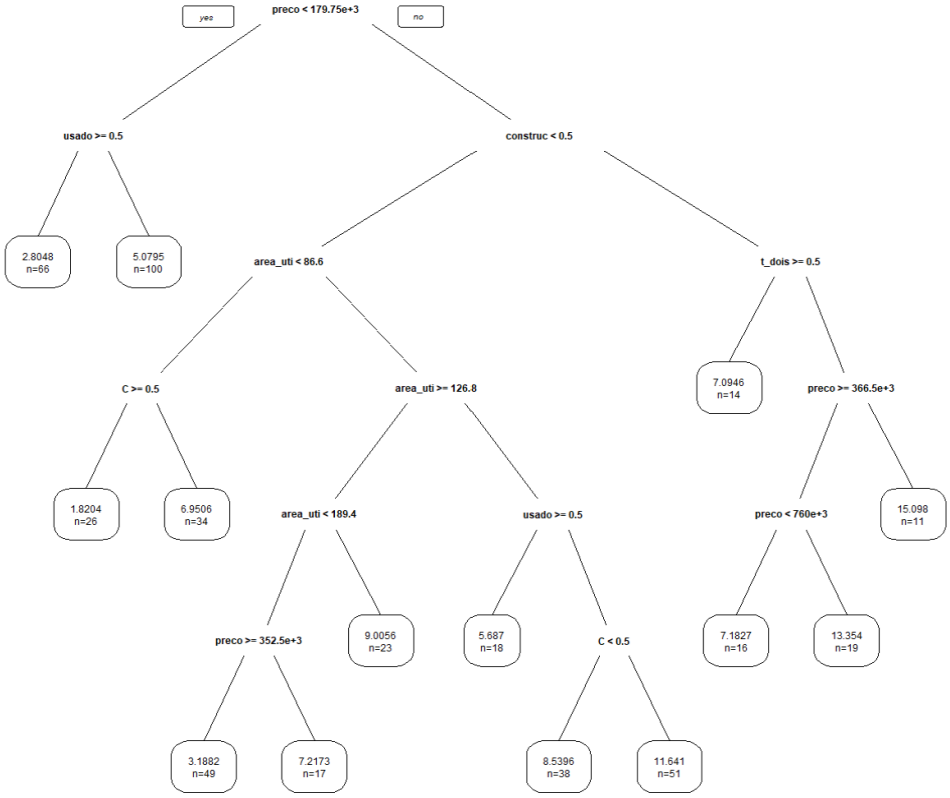


Figura 32 - Árvore de Regressão dos Imóveis do Porto

Relativamente às métricas que permitem avaliar a precisão do modelo, estas melhoraram significativamente, como se verifica na Tabela 3.

Tabela 3 Métricas de Avaliação das Árvores de Decisão da Base de Dados do Porto

	Erro Quadrado Médio	Erro Raiz Quadrada Médio	Erro Médio Absoluto
Árvores de Decisão Porto	11.17	3.34	2.47

Estes valores representam uma melhoria relativamente à base de dados originais, pois os erros diminuíram. Na Figura 33 representa-se o gráfico de *accuracy*, onde a linha reta representa os valores reais, e os círculos vermelhos os valores provenientes da previsão.

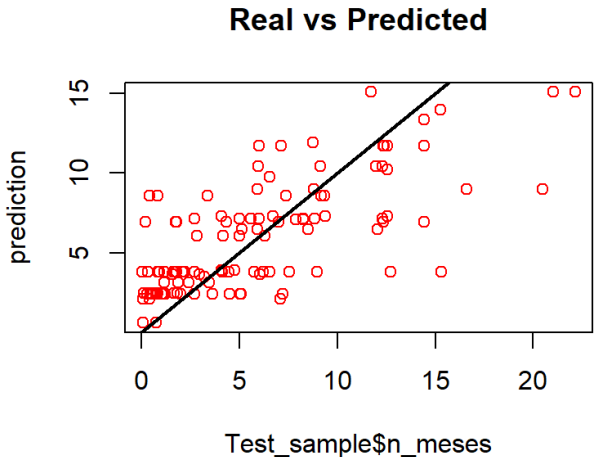


Figura 33 - Gráfico comparativo entre valores reais e previstos

De salientar que para a otimização do método, poder-se-ia recorrer ao *parameter tunning* – método que otimiza os parâmetros do modelo, atingindo-se assim, um melhor desempenho. Este método não foi utilizado no modelo de árvores de regressão para a obtenção do número mínimo de observações num nó, no entanto, o *CP* foi escolhido de acordo com o menor erro.

4.2.2 Redes Neuronais Artificiais

As redes neuronais artificiais pretendem reproduzir o comportamento do cérebro humano, imitando os neurónios na forma de comunicação, e têm como objetivo reconhecer padrões e prever valores em problemas de regressão e classificação.

Estas redes são compostas por camadas de nós, contendo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Os nós são conectados e têm um peso associado; se a saída de qualquer nó individual estiver acima do valor limite especificado, esse nó é ativado, e envia os dados para a próxima camada de rede.

Os pesos associados a cada nó permitem determinar a importância das variáveis de entrada, permitindo determinar quais as mais significativas. Todas as entradas são multiplicadas pelos seus respetivos pesos e somadas. No final, visto que as redes neuronais são complexas, estas escolhem dinamicamente o melhor tipo de regressão para os dados presentes (Bergsma 2013).

As redes neuronais foram construídas recorrendo à função "*neuralnet*", tendo-se arbitrado 5 como valor das camadas ocultas. Posteriormente, foi calculado o erro quadrado médio, e alterado sucessivamente o valor das camadas ocultas para aferir qual o valor que se traduziria num menor erro.

Após um estudo iterativo, para a base de dados pré-processada original, foi encontrada uma solução que minimiza o erro, e que é constituída por 3 camadas ocultas.

À semelhança do efetuado para o método de árvores de decisão, foram comparados os resultados entre a base de dados original e a base de dados com os imóveis apenas da cidade do Porto. No entanto, para a análise e previsão da segunda base de dados, houve a necessidade de eliminar variáveis com pouca relevância para a previsão, pois não existiam observações suficientes para um modelo preditivo eficaz. O critério utilizado foi eliminar as variáveis cuja contribuição era inferior a 2%, ou seja, menor que 10 observações.

Após eliminação e construção da rede neuronal, foi encontrado o valor ótimo de 7 camadas ocultas, que originaram os resultados demonstrados na Tabela 4.

Tabela 4 - Métricas de Avaliação das Redes Neurais

	Erro Quadrado Médio	Erro Raiz Quadrada Médio	Erro Médio Absoluto
Neural Networks original	52.89167	7.272666	4.985774
Neural Networks Porto	3.400862	**1.844143**	1.047529

Em ambos os modelos foi executado um gráfico, com o objetivo de perceber visualmente qual era a diferença entre o valor preditivo e o valor real, à semelhança do efetuado nas árvores de regressão.

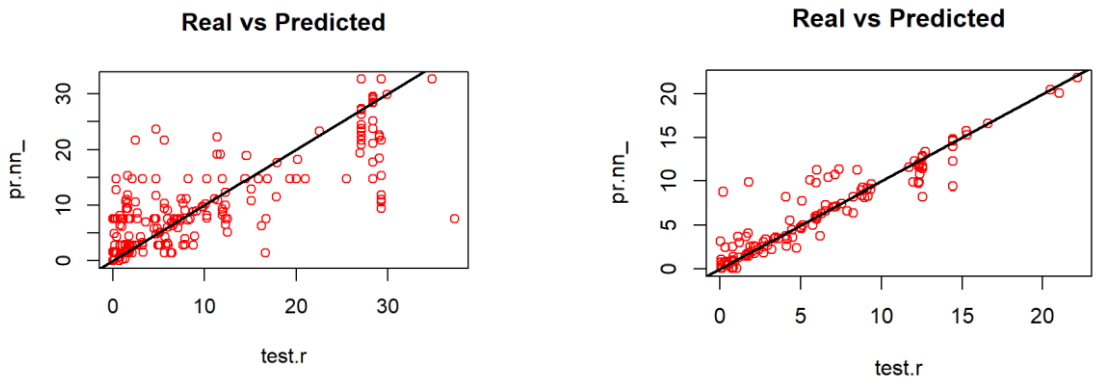


Figura 34 - Gráficos de Comparação entre valor real e preditivo da base de dados original e base de dados do Porto, respetivamente

4.2.3 Nearest Neighbor Regressão

O método *k-NN*, tal como mencionado anteriormente, utiliza a semelhança de observações para prever os valores de novos dados. É, para isso, necessário otimizar um dos parâmetros mais importantes do método: o k (número de vizinhos que são observados até atribuição de valor à nova observação).

Para esta otimização, foi utilizado o *Cross Validation*, um método estatístico de avaliação e comparação de observações que utiliza três partes da base de dados: *training sample*, para encontrar os *nearest neighbors*, *cross validation data* para escolher o melhor valor de *k* e, finalmente, o *testing sample* para testar o modelo com os valores ainda não vistos (Vijay Kotu 2014).

Foram testados valores de *k* de 1 a 15, visto que, a partir desse valor, os erros associados à previsão têm tendência a aumentar. Utilizando como exemplo a base de dados apenas do Porto, onde se obteve um erro mais baixo comparativamente com a original, os valores obtidos para o *k* e respetivos erros foram os apresentados na Figura 35.

k	RMSE	Rsquared	MAE
1	1.0395965	0.2752392	0.6845771
2	0.9062382	0.3155502	0.6283418
3	0.9113044	0.2824427	0.6507387
4	0.9204841	0.2529630	0.6681362
5	0.9315613	0.2275035	0.6748927
6	0.9363068	0.2162830	0.6814560
7	0.9309219	0.2154781	0.6820469
8	0.9220204	0.2217030	0.6778352
9	0.9164987	0.2219942	0.6780291
10	0.9076538	0.2315517	0.6796612
11	0.9067035	0.2288363	0.6798369
12	0.9132692	0.2168842	0.6897005
13	0.9241280	0.1997985	0.7035145
14	0.9320827	0.1860797	0.7148591

A coluna "*Rsquared*" representa a quantidade da variável a prever que é explicada pelo modelo. Assim, o *Root Mean Squared Error* (*RMSE*) é inversamente proporcional ao *Rsquared*.

O gráfico da Figura 35 ilustra todos os pontos testados, para uma melhor compreensão. Neste modelo, o *k* ótimo é de 3 e 2, para as bases de dados original e do Porto, respetivamente.

Após previsão e cálculo de erros, demonstrados na Tabela 5, conclui-se que a base de dados só com os imóveis do Porto obteve erros mais baixos, comparativamente com a original.

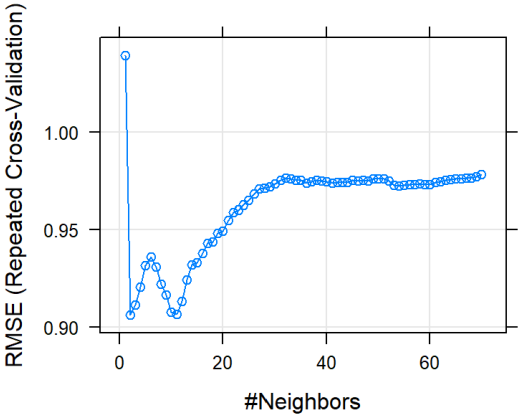


Figura 35 - Valores de k e respetivos erros

Tabela 5 - Métricas de Avaliação do *k-NN*

	Erro Quadrado Médio	Erro Raiz Quadrada Médio	Erro Médio Absoluto
k-NN original	82.62481	9.089819	6.101276
k-NN Porto	23.57096	4.854993	3.268558

4.2.4 Random Forest

O *Random Forest* é um algoritmo baseado no método das árvores de decisão/regressão, embora não haja interação entre elas no momento dos cálculos de previsão. Este algoritmo tem como objetivo reduzir a complexidade dos modelos.

Para esta redução de complexidade, o modelo utiliza o método *bagging*, uma técnica que escolhe uma amostra aleatória do conjunto de dados, e a partir de cada amostra gera um modelo. Cada modelo é treinado de forma independente, e conduz a resultados também eles independentes. O resultado final é baseado na combinação dos resultados de todos os modelos (Breiman 2001).

Este método seleciona, então, observações aleatórias, constrói as árvores de regressão, neste caso, e é calculado o resultado final de acordo com os resultados de cada árvore, não utilizando para isso nenhuma fórmula especifica.

O *random forest* tem alguns parâmetros extremamente importantes, e de difícil definição inicial, como o número de árvores a desenhar e o número inicial de variáveis de previsão a serem consideradas em cada divisão. Assim, e para que estes valores sejam o mais adequados possível, foram utilizados dois métodos: *parameter tuning* e *caret package*.

Os resultados obtidos para a base de dados original foram os apresentados na Figura 36. Através desta, é possível observar que o *caret package*, para um número inicial de variáveis de previsão de 18, conduz a um menor erro, pois é o que atinge um *RMSE* menor.

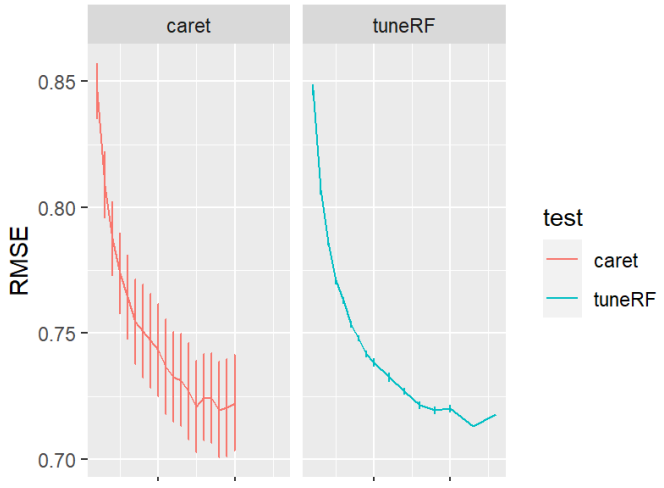


Figura 36 - Comparação do RMSE com *caret* e *parameter tunning*

Relativamente ao número de *regression trees* a utilizar, foi considerado o número de observações da base de dados, testando diferentes valores no intervalo de +/- 200, de 50 em 50.

Na Figura 37 estão representados os resultados obtidos. Cerca de 50% da variância do número de meses é explicada pelas variáveis independentes, e, após previsão, foi possível calcular os erros associados, obtendo-se um *RMSE* de 9,01.

```
Call:
 randomForest(formula = n_meses ~ ., data = Train_sample, ntree = 850,      m
try = 18, keep.forest = TRUE, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 850
No. of variables tried at each split: 18

          Mean of squared residuals: 0.510741
                    % Var explained: 49.3
```

Figura 37 - Resultado do modelo *Random Forest* na base de dados original

Para além dos erros já calculados noutros modelos, no *random forest* foi também possível determinar qual das variáveis contribuiu de forma mais significativa para o aumento do erro quadrado médio, como demonstra a Figura 38.

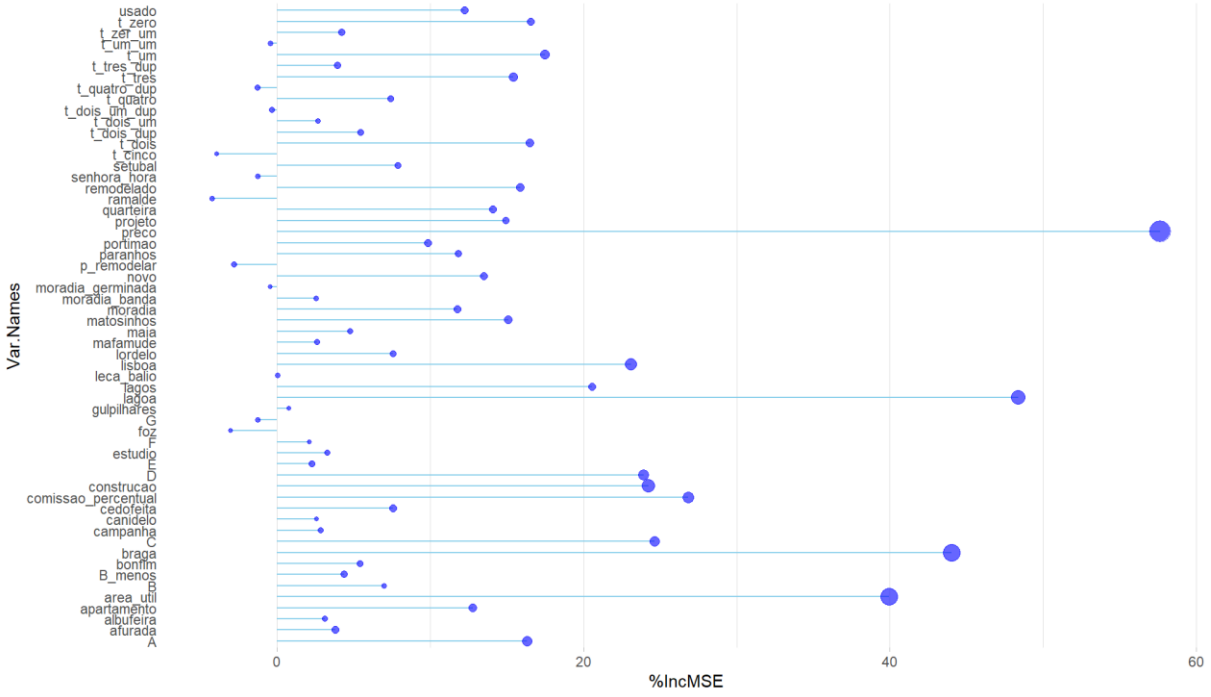


Figura 38 - Aumento de MSE por variável

Tal como esperado, o modelo utilizando a base de dados do Porto obteve melhores resultados, visto que algumas das variáveis que mais aumentam o erro quadrado médio foram eliminadas. Embora tenha sido possível prever o tempo de venda com maior rigor, a percentagem da variável a ser explicada diminuiu para 32%.

Assim, e recorrendo novamente ao *parameter tuning* utilizando o *package caret*, o número ideal de variáveis encontrado é de 13, e foram consideradas 400 árvores de regressão. Os resultados foram os seguintes:

```
Call:
 randomForest(formula = n_meses ~ ., data = Train_sample, ntree = 400,      m
try = 13, keep.forest = TRUE, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 400
No. of variables tried at each split: 13

          Mean of squared residuals: 0.7016903
                    % Var explained: 32.14
```

Figura 39 - Resultados do modelo *Random Forest* para base de dados do Porto

Finalmente, para o erro quadrado médio obteve-se 16,78 meses, o erro raiz quadrada média de 4,10 meses e erro médio absoluto de 3,01 meses.

O modelo de *random forest* apresentou um erro maior face ao obtido no modelo *regression tree*, contrariamente ao esperado, visto que o modelo *random forest* incorpora as soluções de centenas de árvores de regressão, calculando posteriormente a solução final, considerando os resultados das melhores árvores de regressão singulares.

No entanto, uma possível justificação prende-se com uma das desvantagens deste modelo. Este não permite a extrapolação para além do que a base de dados compreende, ou seja, as previsões são sempre realizadas no intervalo do *training sample* (Breiman 2001).

Ambas as bases de dados consideradas contêm variáveis com poucas observações (embora superiores a 5), e o comprimento da base de dados em si não é extenso, dificultando a realização de previsões precisas através de modelos como o *random forest* o *k-NN*.

4.2.5 Resumo de Resultados obtidos

Através dos algoritmos realizados nos subcapítulos anteriores, foi possível chegar aos resultados apresentados na Tabela 6.

Tabela 6 – Erros, em meses, associados aos diferentes modelos

	MSE	RMSE	MAE
	Erro Quadrado Médio	**Erro Raiz Quadrada Média**	Erro Médio Absoluto
Regression Tree original	56.1359	7.492389	5.01304
Regression Tree Porto	11.17233	3.342503	2.468158
Neural Networks original	52.89167	7.272666	4.985774
Neural Networks Porto	3.400862	**1.844143**	1.047529
k-NN original	82.62481	9.089819	6.101276
k-NN Porto	23.57096	4.854993	3.268558
Random Forest original	61.56801	7.846528	5.664665
Random Forest Porto	16.77555	4.095796	3.014422

Conclui-se assim, que através das redes reunorais aplicadas à base de dados do Porto, é atingido um valor ótimo de discrepância, de apenas 1.8 meses. Este valor é o melhor comparando todos os modelos, e traduz um erro baixo de previsão relativamente aos valores reais da *test sample*.

A nível interno, de negócio, este resultado traduzirá uma mais valia quando implementado, pois um erro de 1.8 meses numa variável que normalmente não é prevista, é um bom indicador quer para os clientes proprietários, como para a equipa EURO BROKERS.

4.3 Metodologia de implementação

Para a implementação de uma nova metodologia numa empresa, inicialmente, é necessário definir e explicar de forma clara a todos os envolvidos as alterações necessárias para que esta ocorra. Neste caso, a EURO BROKERS pretende implementar uma métrica adicional que fornecerá ao cliente proprietário, no momento da angariação/consultoria. Para que esta métrica, calculada através de um modelo de *data analytics*, possa ser incorporada nos processos da empresa, será necessário formar todos os colaboradores, para que estes adquiram conhecimentos-base sobre este tema. Adicionalmente, será necessário demonstrar a mais-valia introduzida pela implementação desta métrica para a empresa, bem como para os clientes proprietários.

Prevê-se que a execução e utilização das ferramentas que podem ser desenvolvidas com a informação conseguida ao longo do presente projeto, potenciará uma vantagem competitiva para a EURO BROKERS, permitindo à empresa fornecer um serviço mais completo aos seus clientes.

A previsão do tempo de venda do imóvel permitirá ao cliente proprietário efetuar uma correta gestão dos seus recursos financeiros, visto que a venda de um imóvel implica custos que podem ser minimizados com um planeamento mais informado.

A previsão do tempo de venda do imóvel complementará a análise atual e normalmente realizada pela empresa, e será efetuada assim que a EURO BROKERS possuir as informações relativas às características do imóvel, e antes do contacto presencial com o cliente proprietário.

Na prática, após a introdução dos dados do imóvel, de forma análoga ao realizado atualmente, os colaboradores da EURO BROKERS teriam de correr o modelo de previsão, de forma a obter o tempo de venda estimado, com o respetivo erro associado, tal como mencionado anteriormente (1,8 meses). Adicionalmente, se relevante, será possível analisar a variação deste fator com o preço, comissão percentual, e estado do imóvel.

Esta variação é importante para o proprietário perceber de que forma pode diminuir o tempo de venda, por exemplo, se remodelar o imóvel, se aumentar a comissão percentual, entre outros. De salientar que esta dimensão tem ainda de ser trabalhada, cujas dimensões serão aprofundadas no capítulo seguinte.

5 Conclusões e perspetivas de trabalho futuro

O presente projeto consistiu na previsão do tempo de venda de um imóvel recorrendo a soluções de *data analytics*. No final do projeto foi cumprido o cronograma definido inicialmente e o objetivo principal do mesmo foi atingido. Para além do desenvolvimento de um modelo de previsão do tempo de venda do imóvel, o presente projeto permitiu dar a conhecer o potencial de soluções de *data analytics*, presentes no mercado ou a desenvolver, que poderão ser uma mais-valia para a atividade da EURO BROKERS.

A construção dos algoritmos para os modelos de previsão foi executada. No entanto, até ao momento, ainda não foi possível os mesmos serem testados pela equipa da EURO BROKERS, embora seja um objetivo num futuro muito próximo. Para além da análise profunda aos imóveis que constituem a carteira da empresa, foram também analisadas informações relativas aos clientes compradores e proprietários.

Relativamente aos clientes compradores, é possível concluir que, nos anos 2019, 2020 e 2021, o número de clientes a entrarem em contacto com a EURO BROKERS manteve-se estável.

Cerca de 60% dos contactos foram direcionados para compra, enquanto que apenas 32% para arrendamento. Relativamente ao tipo de clientes compradores, 75% do total foram classificados pela EURO BROKERS como potenciais clientes, o que possibilita à empresa contactá-los quando surgir um imóvel com as características requeridas por estes.

Os clientes proprietários são na sua maioria do género masculino, e em média cada proprietário possuiu 1,28 $\approx$ 1 imóvel. O concelho do Porto é o que reúne mais clientes, seguido de Matosinhos e Vila Nova de Gaia.

Um dado importante relativamente a estes clientes é o facto do número de novos proprietários ter vindo a diminuir ao longo dos anos, no entanto a empresa manteve-se em lucro. É objetivo aumentar o número de imóveis em carteira, para de uma forma consistente crescer a empresa e chegar a cada vez mais clientes.

A base de dados dos imóveis foi a mais desafiante de construir, uma vez que esta continha uma grande quantidade de informação irrelevante. Esta etapa exigiu um esforço considerável que se traduziu na construção dos algoritmos para os modelos de previsão, dos quais se obtiveram os resultados relativos aos erros associados a cada um dos modelos (Tabela 4 do Capítulo 4).

Relativamente aos modelos de previsão realizados, estes poderiam ter sido mais otimizados recorrendo a técnicas de *data analytics* mais avançadas, no entanto, este processo constituiu um grande obstáculo devido à falta de experiência e conhecimento. De salientar, que todo o conhecimento relativamente a este tópico foi adquirido em apenas 4 meses, em consolidação com a Unidade Curricular de Analítica Empresarial, cujas análises incidiram maioritariamente em problemas de classificação.

Posto isto, e de acordo com a análise realizada, o método *Neural Networks* aplicado à base de dados dos imóveis do Porto é o que representa um erro quadrático médio menor, de 1,8

meses, ou seja, prevê-se que a venda do imóvel aconteça 1,8 meses acima ou abaixo do valor estimado.

A métrica de previsão desenvolvida ao longo deste projeto necessita de otimizações no futuro, para que se torne progressivamente mais precisa e exata, nomeadamente envolvendo uma recolha e tratamento de dados mais amplos, aumentando, assim, o leque de características associadas a cada imóvel e o número das respetivas observações.

Para além disso, no presente projeto, a base de dados utilizada possuía apenas 1000 observações, tendo sido reduzida para metade na base de dados relativa ao distrito do Porto, sendo as variáveis presentes insuficientes para prever a variância do tempo de venda de um imóvel. Assim, sugere-se que, no futuro, seja realizada uma recolha de dados mais robusta e adequada, de forma a aprimorar o modelo de previsão.

Adicionalmente, fatores externos, tais como a existência de concorrência, situação económica, ciclo do mercado, e características da localização (proximidade de infraestruturas, tais como parques, escolas ou farmácias), devido à potencial influência no tempo de venda de um imóvel, deverão ser considerados em futuras iterações do modelo de previsão desenvolvido.

Todos os fatores mencionados permitirão à EURO BROKERS alavancar os conhecimentos gerados ao longo do presente projeto, potenciando um maior crescimento da sua atividade, alicerçado em modelos de avaliação automatizada, bem como em soluções para o apoio à tomada de decisão e índices de preço de propriedades.

Referências

Benesty, Jacob, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. "Pearson correlation coefficient." In Noise reduction in speech processing, 1-4. Springer.

Bergsma, Wicher. 2013. "A bias-correction for Cramér's V and Tschuprow's T." Journal of the Korean Statistical Society 42 (3): 323-328.

Bishop, Chris M. 1994. "Neural networks and their applications." Review of scientific instruments 65 (6): 1803-1832.

Breiman, Leo. 2001. "Random forests." Machine learning 45 (1): 5-32.

Cattaneo, Laura, Luca Fumagalli, Marco Macchi, and Elisa Negri. 2018. "Clarifying data analytics concepts for industrial engineering." IFAC-PapersOnLine 51 (11): 820-825.

Cheryshenko, MS, and Yu Yu Pomernyuk. 2021. "Integration of big data in the decision-making process in the real estate sector." IOP Conference Series: Earth and Environmental Science.

Damin, Hiram. 2019. Customer Success: O sucesso das empresas focadas em clientes. DVS Editora.

De Mauro, Andrea, Marco Greco, and Michele Grimaldi. 2016. "A formal definition of Big Data based on its essential features." Library review.

Fawcett, Foster Provost and Tom. Data Science for Business. 1 ed. edited by Mike Loukides and Meghan Blanchette, July 2013.

Fayyad, Usama M, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework." KDD.

Huang, Jianglin, Jacky Wai Keung, Federica Sarro, Yan-Fu Li, Yuen-Tak Yu, WK Chan, and Hongyi Sun. 2017. "Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study." Journal of Systems and Software 132: 226-252.

Loh, Wei-Yin. 2011. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1 (1): 14-23.

Park, Sun Jung Park. 2020. "Data science strategies for real estate development." Massachusetts Institute of Technology.

Reis, Edna Afonso, and Ilka Afonso Reis. 2002. "Análise descritiva de dados." Relatório Técnico do Departamento de Estatística da UFMG 1.

Rodrigues, João Manuel Pereira. 2008. "Especificação de um modelo Hedónico de preços de habitação para Portugal." Instituto Superior de Economia e Gestão.

Rohit Amarnath. 2022. "Council Post: Eight Trends Predicted to Define Data Analytics in 2022." Forbes, April 14, 2022. Disponível em:

https://www.forbes.com/sites/forbestechcouncil/2022/02/25/eight-trends-predicted-to-define-data-analytics-in-2022/?sh=172ddef0ffd7.

Runkler, Thomas A. Data Analytics - Models and Algorithms for Intelligent Data Analytics. 3 ed., edited by Vieweg Teubner Verlag. 28 setembro 2012. doi:10.1007/978-3-658-29779-4.

Tsai, Chun-Wei, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V Vasilakos. 2015. "Big data analytics: a survey." Journal of Big data 2 (1): 1-32.

Unacast.com. 2021. “The Top Real Estate Analytics Companies Right Now.” 2021. https://www.unacast.com/post/the-top-real-estate-analytics-companies-right-now

Vijay Kotu, Bala Deshpande. Predictive Analytics and Data Mining. 1 ed., 2014.

Winson-Geideman, Kimberly, and Andy Krause. 2016. "Transformations in real estate research: The big data revolution." Proceedings of the 22nd Annual Pacific-Rim Real Estate Society Conference, Queensland, Australia. Disponível em: http://www.prres.net/papers/Geideman_Transformations_in_RE_Research.pdf

ANEXO A: Resultados das Análises

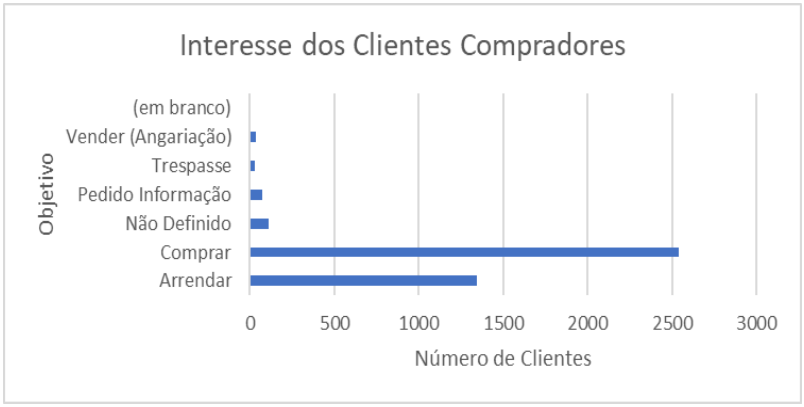


Figura 41 - Número de Clientes Compradores por Objetivo de negócio

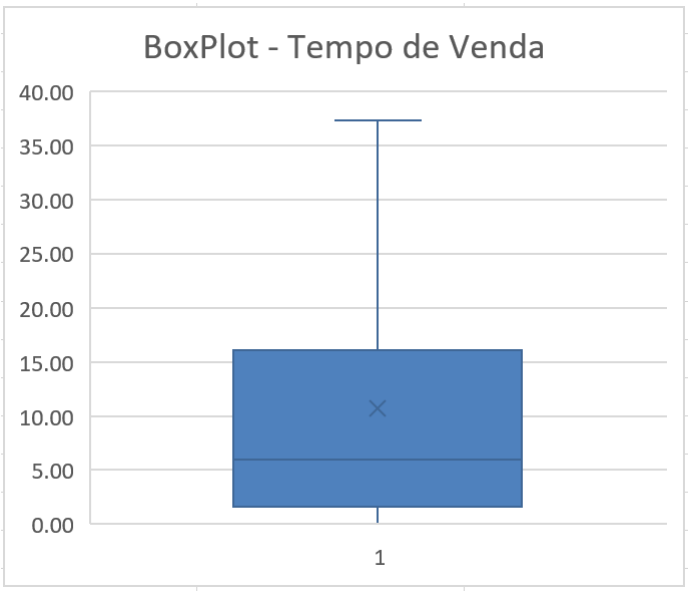


Figura 42 – BoxPlot do Tempo de Venda da Base de Dados Original

Condição 1	Condição 2	Condição 3	Condição 4	Condição 5	Condição 6	Condição 7	Condição 8	Condição 9	Condição 10	Condição 11	Condição 12	Condição 13	Condição 14	Valor	Número Imóveis	Tempo de Venda (meses
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1	ñ T3	area útil < 116.82	ñ usado	preço < 277200	preço<	234064.5616	72	14.39707498
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1	ñ T3	area útil < 116.82	ñ usado	preço < 277200	preço>=	234064.5616	26	21.65567248
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1	ñ T3	area útil < 116.82	ñ usado		preço >=	277200	31	12.99
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1	ñ T3	area útil < 116.82			usado		7	7.277781437
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75	ñ usado	ñ C	ñ A				preço<	477432.6484	21	5.392338708
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75	ñ usado	ñ C	ñ A				preço >=	477432.6484	23	9.676877375
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1	ñ T3				area_util>	116.797717	17	20.72073622
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita	ñ T1					T3		9	9.564914251
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75	ñ usado	ñ C					classe A		7	13.14725105
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção	ñ cedofeita						T1		11	24.92593275
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75	ñ usado						classe C		55	11.42469494
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil < 85.75	ñ C						novo		13	6.286718841
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil < 85.75	ñ C						não novo		10	12.33103406
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil < 85.75							classe C		20	1.370432128
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75							usado		8	4.677611689
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil < 143.95	comissão > 0.029	area útil > 85.75							não usado		106	9.964115819
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil > 129.33	area útil < 136.15							Sem ser A		19	6.464682207
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil > 129.33	area útil < 136.15							A		7	17.16517583
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33	ñ construção							cedofeita		12	6.710724981
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D	preço<160750	ñ Cedofeita							area_util <	39.67452906	21	1.60723988
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D	preço<160750	ñ Cedofeita							area_util >	84.50022612	30	2.330369642
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D	preço<160750	ñ Cedofeita							area_util <	77.72889682	34	4.393901674
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D	preço<160750	ñ Cedofeita							area_util >	77.72889682	10	9.818015526
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D	preço<160750								cedofeita		15	8.27177782
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço < 217000	preço > 192500								preço<	199533.2965	14	2.1433848
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço < 217000	preço > 192500								preço>=	199533.2965	36	6.69578008
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço < 217000	preço < 192500								sem ser em lagos		10	5.585389339
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço < 217000	preço < 192500								lagos		12	25.07331202
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil > 129.33								area_util <	136.1472278	7	1.916635589
ñ braga	preço < 777500	preço > 179500	comissão > 0.045	preço > 217000	area útil < 129.33								contrução		22	7.451674296
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D									preço >=	169599.5422	33	5.961066294
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026	ñ D									preco <	169599.5422	31	11.03866048
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil > 143.95									area_util <	193.2412912	38	3.240468916
ñ braga	preço < 777500	preço > 179500	comissão < 0.045	area útil > 143.95									area_util >=	193.2412912	9	10.45724741
ñ braga	preço > 777500	area útil > 93.965	area útil < 166.26	preço < 1027000									moradia		8	21.95
ñ braga	preço > 777500	area útil > 93.965	area útil < 166.26	preço < 1027000									não moradia		13	11.74
ñ braga	preço < 777500	preço < 179500	comissão >= 0.026										classe D		83	2.858861966
ñ braga	preço > 777500	area útil > 93.965	area útil > 166.26										T3		13	7.118823623
ñ braga	preço > 777500	area útil > 93.965	area útil > 166.26										não_T3		21	13.00264021
ñ braga	preço > 777500	area útil > 93.965	area útil < 166.26										preço>=	1027134.499	12	24.09031268
ñ braga	preço < 777500	preço < 179500											comissão <	0.026482676	17	19.2203106
ñ braga	preço > 777500												area_util<	93.9600981	29	28.35822262
*													braga		30	32.60623448

Figura 43 - Resultados da Árvore de Regressão da Base de Dados Original

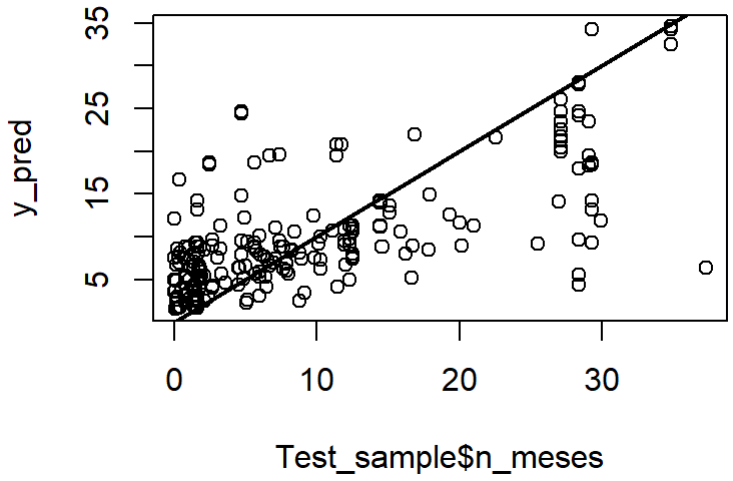


Figura 44 - Gráfico comparativo entre valores reais e previstos da base de dados original, no modelo Random Forest

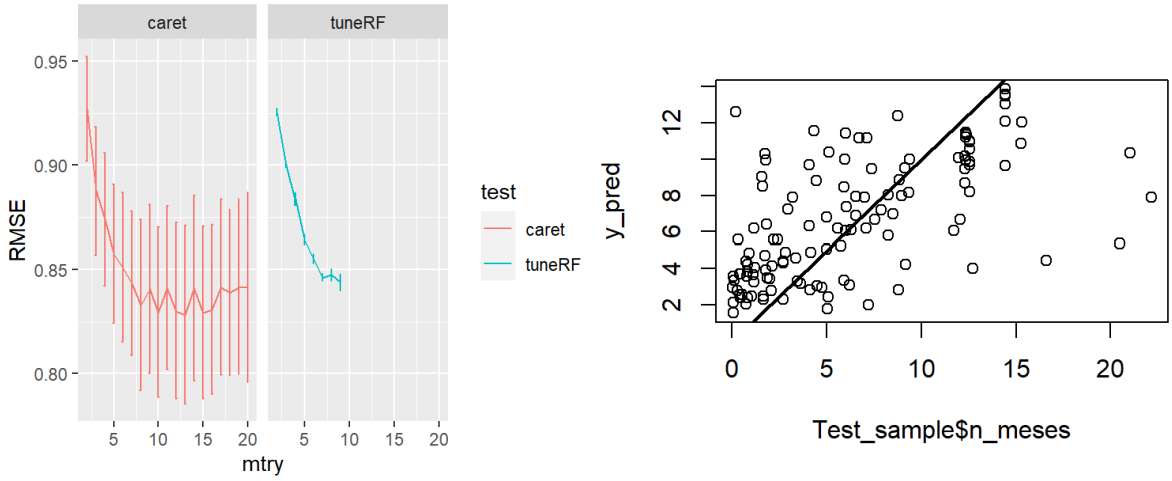


Figura 45 - Comparação do RMSE com caret e parameter tunning e dados reais e previsto, respetivamente, da base de dados do Porto, no modelo Random Forest

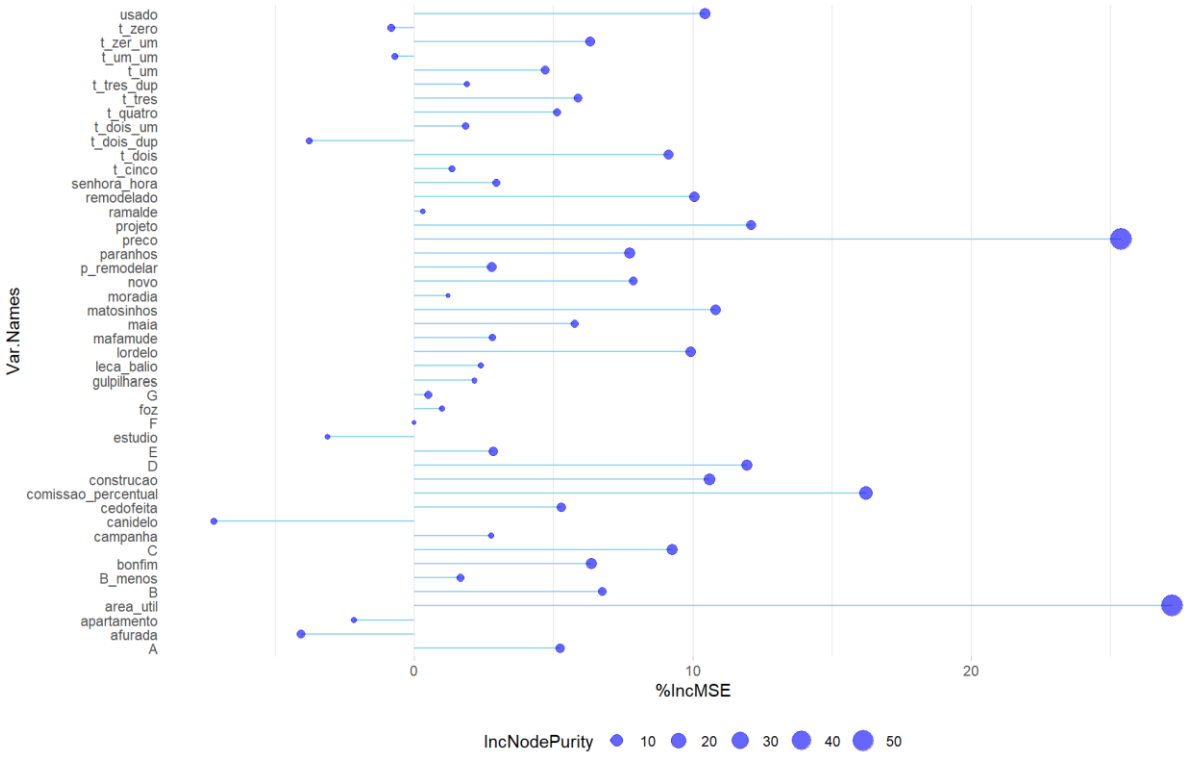


Figura 46 - Aumento de MSE por variável da base de dados do Porto

Variáveis				
area_util	t_dois	campanha	quarteira	D
preco	t_dois_dup	canidelo	ramalde	E
comissao_percentual	t_dois_um	cedofeita	rio_tinto	F
n_meses	t_dois_um_dup	ermesinde	afurada	G
apartamento	t_dois_dois	foz	santo_tirso	
estudio	t_tres	grijo	senhora_hora	
herdade	t_tres_dup	gulpilhares	setubal	
moradia	t_tres_tri	lagoa	valongo	
moradia_devoluta	t_tres_um	lavra	vila_conde	
moradia_banda	t_tres_um_dup	leca_balio	vilar_andorinho	
moradia_germinada	t_quatro	lisboa	construcao	
moradia_isolada	t_quatro_dup	lordelo	projeto	
moradia_restaurar	t_quatro_tri	madalena	novo	
predio	t_quatro_um	mafamude	p_remodelar	
t_zero	t_cinco	maia	remodelado	
t_zero_dup	t_cinco_dup	matosinhos	usado	
t_zer_um	t_seis	oliveira_douro	A	
t_um	albufeira	paranhos	A_mais	
t_um_dup	aveiro	pedroso	B	
t_um_um	bonfim	portimao	B_menos	
t_um_dois	braga	povoa_varzim	C	

Figura 47 - Variáveis após transformação em *dummy variables*