

Previsão do Potencial de Vendas na Cadeia de Pescado com base na Procura Censurada

Mariana Teixeira de Jesus

Dissertação de Mestrado

Orientador na FEUP: Prof. Vera Lucia Miguéis Oliveira e Silva

Coorientador na FEUP: Prof. Luís Gonçalo Rodrigues Reis Figueira



Mestrado Integrado em Engenharia e Gestão Industrial

2020-06-29

Resumo

Portugal é o país da União Europeia com o maior consumo de peixe fresco, *per capita*. A localização geográfica do país e o contexto histórico-cultural justificam este consumo. A natureza perecível do peixe torna fundamental um elevado conhecimento dos consumidores e um sistema de aprovisionamento capaz de responder às necessidades dos retalhistas. Um dos grandes desafios dos atuais modelos de previsão de vendas é não considerarem a procura censurada, isto é, casos em que as vendas registadas não refletem a procura dos clientes. Porquanto os modelos são elaborados com os valores anteriores de vendas, pode existir uma subestimação da procura, o que consequentemente provoca erros nos modelos futuros.

Este projeto procura obter uma perceção correta dos padrões de consumo dos clientes de peixe fresco. Primeiramente, é importante perceber como o mercado do pescado tem evoluído ao longo dos anos analisados. Num passo seguinte, é fundamental analisar os dados fornecidos pela empresa de retalho e perceber em que situações se verifica censura da procura. Para estes casos, é necessário criar heurísticas que permitam resolver a subestimação dos valores e demonstrem a procura real. A heurística que apresenta melhor desempenho é baseada nas vendas observadas no dia da semana homólogo, ou seja, a média entre a quantidade vendida no mesmo dia da semana anterior e seguinte.

Num passo seguinte, é necessário a elaboração de modelos de previsão de vendas para um conjunto de dados de teste, tendo como base a nova procura calculada. Realizaram-se cinco modelos, três modelos de *Machine Learning* de natureza causal: Rede Neural Artificial, *Support Vector Regression* e *Random Forest* e dois modelos estatísticos *time-series*: SARIMA e *Holt Winters*. A implementação do método RF conduz a melhores resultados, conseguindo ultrapassar o desempenho do modelo atualmente utilizado pela empresa de retalho. Os modelos estatísticos geram resultados com desempenho pior em comparação com os modelos de ML.

Por último, é fundamental comparar os resultados dos modelos com a correção dos valores censurados da procura e quando essa correção não acontece. Na generalidade, quando se lida com a procura censurada a quantidade prevista pelos modelos de previsão de vendas é superior às situações em que não se lida. Consequentemente, a transformação da quantidade vendida é fundamental para evitar a subestimação da procura.

Abstract

Potential Sales Forecasting in the Fish Chain based on Censored Demand

Portugal is the country of the European Union with the highest consumption of fresh fish, *per capita*. The country's geographical location and the historical-cultural context justify this consumption. The perishable nature of fish makes it essential to have a high level of consumer knowledge and a supply system capable of meeting the needs of retailers. One of the major challenges of current sales forecasting models is not to consider censored demand, that is, cases in which registered sales do not reflect customer demand. As the models use previous sales data, there may be an underestimation of demand, which consequently causes errors in future models.

The main goal of this project is to obtain a correct perception of the consumption patterns of fresh fish customers. First, it's important to understand how the fish market has evolved over the years. In a next step, it's essential to analyze the data provided by the retail company and understand in which situations there is demand censorship. For these cases, it's necessary to create heuristics that allow to resolve the underestimation of the values and estimate the potential real demand. The heuristic with the best performance is based on the sales of the same day of the week, that is, the average between the quantity sold on the same day of the previous and the following week.

In a next step, it's necessary to develop sales forecast models based on the new calculated demand. Five models were made, three machine learning models with explanatory variables: Artificial Neural Networks, Support Vector Regression and Random Forest and two statistical time-series models: SARIMA and Holt Winters. The implementation of the RF method leads to better results, managing to outperform the model currently used by the retail company. Statistical models generate results with worse performance compared to ML models.

Finally, it's essential to compare the results of the models with the correction of the censored values of demand and when this correction does not happen. In general, when dealing with censored demand, the quantity predicted by the sales forecast models is higher than the situations in which it's not dealt with. Consequently, the transformation of the quantity sold is essential to avoid underestimating of demand.

Agradecimentos

Após cinco anos a estudar Engenharia e Gestão Industrial, esta dissertação de mestrado representa o corolário de uma caminhada, da qual eu não poderia estar mais grata e satisfeita.

Primeiramente, gostaria de expressar o meu agradecimento à minha orientadora, Professora Vera Miguéis, pela eterna disponibilidade e diligência em responder às minhas dúvidas e inquietações. Não podendo deixar de agradecer ao Professor Gonçalo Figueira pela sua cooperação.

Gostaria também de gratificar a empresa retalhista parceira do projeto, Sonae MC, pela disponibilidade na disponibilização dos dados necessários para a concretização deste trabalho.

Sem a minha família este percurso não seria possível, aos meus Pais e Irmã, pelo apoio durante todo este percurso. Obrigada por estarem sempre presentes, mesmo quando as horas em casa eram escassas, pela motivação e por todas as pequenas coisas que fizeram e muito impactaram esta minha jornada.

Aos amigos, os de sempre, que não cobram tempo nem disponibilidade, que me incentivam a ser mais e nunca deixar de sonhar. A todos aqueles que partilharam comigo muitos momentos ao longo desta aventura, quer na L203 ou na máquina de café.

Finalmente, ao grupo de pessoas que me acompanhou durante estes cinco anos e que tornou esta aventura tão especial. Aquelas que muito me fizeram rir, chorar e que fizeram estes anos valerem a pena. Nunca esquecerei os dias e noites de estudo, todos os fins de semana juntos, os jantares e festas, as discussões e as ideias que pareciam estúpidas, mas que se revelaram incríveis. Que nunca deixemos de ser irreverentes juntos.

Um obrigada a estes e a todas as pessoas que tornaram estes anos tão especiais e que irei sempre recordar com muito carinho.

Mariana Teixeira de Jesus

"There is no passion to be found playing small - in settling for a life that is less than the one you are capable of living."

Nelson Mandela

Conteúdo

1	Introdução	1
1.1	Enquadramento do projeto	1
1.2	Motivação	2
1.3	Objetivos e metodologia	2
1.4	Estrutura da dissertação	3
2	Revisão de literatura	5
2.1	Comportamento dos Consumidores de Peixe Fresco	5
2.2	Previsão de Vendas no Setor do Retalho	6
2.3	Previsão de Vendas	8
3	Caracterização das Vendas de Peixe Fresco	11
3.1	Descrição da Empresa	11
3.2	Caracterização dos Dados	12
3.3	Conclusão	15
4	Metodologia	17
4.1	Processo de <i>Data Mining</i>	17
4.1.1	Seleção	19
4.1.2	Pré-processamento	21
4.1.3	Transformação	21
4.1.4	Técnicas de <i>Data-Mining</i>	22
4.1.5	Algoritmos de <i>Data Mining</i>	23
4.1.6	Validação dos Modelos	27
4.1.7	Tecnologias Utilizadas no Projeto	28
5	Resultados	29
5.1	Heurísticas para a Procura Censurada	29
5.2	Modelos de Previsão de Vendas	30
5.2.1	Dourada Média	31
5.2.2	Salmão Inteiro	32
5.2.3	Robalo Médio	32
5.3	Impacto da Procura Censurada	33
5.4	Importância das Variáveis	35
6	Conclusão e Trabalho Futuro	39
A	Caraterização das Variáveis	43

Acronyms and Symbols

ML	<i>Machine Learning</i>
ANN	Rede Neural Artificial
ARIMA	Modelo Auto-regressivo Integrado de Médias Móveis
SKU	<i>Stock Keeping Unit</i>
KDD	<i>Knowledge Discovery in Databases</i>
SVR	<i>Support Vector Regression</i>
SVM	<i>Support Vector Machine</i>
RF	<i>Random Forest</i>
SARIMA	Modelo Auto-regressivo Integrado de Médias Móveis Sazonal
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
RMSE	Raiz Quadrada do Erro Médio
HW	<i>HoltWinters</i>
F	Frequência
DP	Desvio Padrão

Lista de Figuras

3.1	Vendas Totais de Peixe Fresco por Loja	13
3.2	<i>Treemap</i> das Vendas de Peixe por <i>SKU</i> e categoria	14
3.3	Vendas de Dourada Média por Trimestre	14
3.4	Vendas de Dourada Média por Semana	15
3.5	Vendas Acumuladas de Dourada Média por Dia	15
4.1	Visão geral dos processos de <i>Data Mining</i> (Steiner and Neto, 2006)	18
4.2	Técnicas de <i>Data Mining</i>	23
4.3	Funcionamento de um <i>Perceptron</i> (Deshpande, 2017)	24
4.4	Representação de uma Rede Neural Artificial (Miller, 2020)	24
4.5	Representação de um hiperplano criado pelo método SVM (Silva and Santiago, 2017)	25
4.6	Representação de um modelo de <i>Random Forest</i> (Machado and Corbellini, 2015)	26
5.1	Comparação da Quantidade Prevista de Dourada Média Com e Sem o Tratamento da Procura Censurada	33
5.2	Comparação da Quantidade Prevista de Salmão Inteiro Com e Sem o Tratamento da Procura Censurada	34
5.3	Comparação da Quantidade Prevista de Robalo Médio Com e Sem o Tratamento da Procura Censurada	34
5.4	Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Dourada Média	36
5.5	Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Salmão Inteiro	36
5.6	Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Robalo Médio	37

Lista de Tabelas

5.1	Resultados das Heurísticas para as Vendas Censuradas de Dourada Média	30
5.2	Resultados das Heurísticas para as Vendas Censuradas de Salmão Inteiro	30
5.3	Resultados das Heurísticas para as Vendas Censuradas de Robalo Médio	30
5.4	Resultados dos Modelos de Previsão de Vendas de Dourada Média	31
5.5	Resultados dos Modelos de Previsão de Vendas de Salmão Inteiro	32
5.6	Resultados dos Modelos de Previsão de Vendas de Robalo Médio	33
5.7	Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Dourada Média	35
5.8	Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Salmão Inteiro	35
5.9	Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Robalo Médio	35
A.1	Variáveis presentes nos modelos de Dourada Média	43
A.2	Variáveis presentes nos modelos de Salmão Inteiro	44
A.3	Variáveis presentes nos modelos de Robalo Médio	44

Capítulo 1

Introdução

O presente trabalho foi realizado no âmbito do projeto de dissertação do Mestrado Integrado de Engenharia e Gestão Industrial.

1.1 Enquadramento do projeto

O Projeto VALORMAR - Valorização Integral dos Recursos Marinhos: Potencial, Inovação Tecnológica e Novas Aplicações trata-se de um consórcio que envolve 38 entidades com diversas proveniências, com o objetivo de desenvolver soluções tecnológicas e inovadoras de forma a potencializar a valorização dos recursos marinhos portugueses. Uma das vertentes é o estudo aprofundado do consumidor e do seu comportamento no que concerne à compra de peixe fresco, onde a parceria pré-estabelecida com uma empresa de retalho portuguesa tem um papel preponderante.

O comportamento dos clientes, refletido na procura de peixe, tem merecido a atenção por parte deste retalhista. É a partir da estimativa da procura que a empresa gere o seu aprovisionamento, à qual está inerente a probabilidade de uma situação de excesso de inventário ou de rotura.

Atualmente, os modelos de previsão de vendas utilizados pela empresa não têm em conta o facto do histórico de vendas refletir uma procura censurada, em que a quantidade vendida pode ser diferente da procura, nomeadamente em casos de falta de inventário. Ilustrando, um cliente pode deslocar-se a uma loja para comprar peixe e já não haver inventário que permita satisfazer a sua necessidade. Neste caso, as vendas registadas não irão refletir a procura deste cliente. Portanto, uma previsão de procura baseada no histórico de vendas pode levar a uma subestimação da procura.

Para as empresas de retalho, um bom sistema de previsão de vendas é fundamental, para uma melhor gestão das estratégias e consequentemente atingir metas com maior facilidade. Assim, o nível de serviço apresentado aos clientes melhora já que é possível evitar *stock-outs* e reduzir o desperdício.

1.2 Motivação

Segundo a Organização das Nações Unidas para a Alimentação e Agricultura (FAO, 2018), o consumo global de peixe *per capita* duplicou desde 1960. Portugal é o país com maior consumo de peixe na União Europeia e, a nível mundial só fica atrás da Islândia e do Japão (Expresso, 2017). O consumo anual *per capita* de peixe em Portugal é mais do dobro da média da União Europeia, 55,6 kg e 22,3 kg, respetivamente.

Existem vários fatores que justificam a atração pelo consumo de peixe em Portugal, nomeadamente a sua localização geográfica, com uma área costeira de 1 214 km, e o seu contexto histórico-cultural uma vez que a pesca é uma atividade económica que se sobressaiu historicamente muito cedo. Os últimos dados da FAO (2017) sustentam a importância que a pesca tem no país, mostrando que 179 758 toneladas de peixe em Portugal eram capturadas e apenas 9 800 toneladas eram provenientes de aquacultura. Segundo o mesmo relatório da FAO, 17 536 pessoas encontram-se empregadas no setor da pesca em Portugal mas muitas outras têm também um papel fundamental em áreas relacionadas: transporte, armazenamento, venda e investigação.

Com o livre acesso à informação e a facilidade de a encontrar, cada vez mais os consumidores tomam decisões mais ponderadas, valorizando a sustentabilidade e a origem portuguesa do produto. Desta forma, cresce uma maior responsabilidade por parte das indústrias alimentares para satisfazer as exigências dos consumidores da atualidade. Por outro lado, sendo também o mercado do retalho alimentar extremamente competitivo, é de particular interesse o estudo das áreas onde o mercado do pescado é mais sensível, sobretudo quanto à qualidade e frescura do produto. Portanto, é essencial entender plenamente os padrões de consumo dos portugueses para as empresas conseguirem responder de uma forma expedita aos mesmos.

1.3 Objetivos e metodologia

O projeto visa obter um melhor entendimento dos padrões de consumo dos clientes de peixe fresco. Primeiramente, é importante perceber como o mercado do retalho, em especial do pescado, tem evoluído ao longo dos anos. Os dados fornecidos por uma empresa de retalho possibilita analisar estes padrões de consumo e perspetivar as vendas para o futuro.

Através dos valores históricos de vendas e também de inventário, o projeto pretende desenvolver um modelo de previsão de vendas que considere os momentos de rotura de *stock* como momentos em que as vendas registadas nos sistemas de informação representam uma procura censurada. Assim, será necessário fazer uma inferência da potencial procura, denominada procura real, a partir dos seus valores censurados, testando para isso várias heurísticas. Depois de inferida a procura real, prever-se-á os valores futuros da procura para assim apoiar o retalhista nas suas operações. O projeto tem ainda como objetivo avaliar o impacto da procura censurada registada pela empresa nos modelos de previsão de procura futura, recorrendo assim à comparação dos valores previstos de vendas para o caso em que a procura censurada é e não é considerada.

1.4 Estrutura da dissertação

Além deste capítulo introdutório, o relatório de tese apresenta mais cinco capítulos.

No Capítulo 2, Revisão de Literatura, são analisados trabalhos anteriores nas áreas do comportamento do consumidor de peixe fresco, bem como dos modelos de previsão de vendas, especialmente na área do retalho e de formas de lidar com a procura censurada.

O Capítulo 3, Caracterização das Vendas de Peixe Fresco, destina-se à apresentação do problema e de todas as variáveis que distinguem e influenciam o mercado do pescado e em particular a empresa de retalho.

O Capítulo 4, Metodologia, ocupa-se pela descrição do trabalho desenvolvido, com a explicação dos métodos utilizados e da sua aplicação no contexto do problema.

No Capítulo 5, Resultados, são apresentados os modelos e as heurísticas implementadas com os respetivos resultados e a análise de desempenho.

No Capítulo 6, Conclusão e Trabalho Futuro, sumariza-se o trabalho realizado, sublinhando-se as principais descobertas e resultados bem como potenciais pontos de partida para a continuação do trabalho.

Capítulo 2

Revisão de literatura

Neste capítulo são analisados trabalhos existentes na área e que permitem definir o estado de arte. Primeiramente, é realizado um breve enquadramento ao Comportamento dos Consumidores de Peixe Fresco, sendo fundamental para perceber as suas escolhas e aprimorar os modelos de previsão de vendas. A previsão de vendas no setor do retalho é também explorada neste capítulo, bem como a procura censurada. Por último, realiza-se a análise das previsões de vendas, de uma forma mais macroscópica.

2.1 Comportamento dos Consumidores de Peixe Fresco

Tal como em Portugal, a indústria da pesca tem um papel relevante na economia de vários países da União Europeia. Em Espanha, Claret and Rodríguez-Rodríguez (2012) identificaram as características mais importantes para os consumidores espanhóis na hora de escolher o peixe. Os fatores avaliados por este estudo espanhol são o "país de origem" (Espanha, Marrocos ou Noruega), o "método de captura" (peixe capturado do mar ou produzido em aquacultura), as "condições de armazenamento" (refrigerado ou congelado) e o "preço de compra" (6€/kg, 12€/kg ou 18€/kg). O estudo conclui que o fator mais importante na escolha do peixe para os inquiridos é o "país de origem", mesmo quando este fator é muitas vezes influenciado por outras variáveis, nomeadamente a perceção da frescura do peixe. Os restantes três fatores apresentam uma importância similar moderada para os consumidores na hora da escolha. De realçar que a preferência da população espanhola inquirida incide, na globalidade dos fatores, sobre o peixe fresco selvagem, de origem nacional e vendido entre os 6 e os 12€ por quilograma.

Nos últimos anos, vários estudos têm sugerido que o consumo de peixe fresco encontra-se relacionado com inúmeros benefícios para a saúde, não só para o desenvolvimento de fetos e crianças como também para o combate de doenças cardiovasculares nos adultos (IOM, 2017). A perceção pelo consumidor dos benefícios do peixe para a saúde modela os hábitos de consumo de peixe, o que foi sustentado pelo estudo de Bemrah and Volatier (2008) que conclui que, independentemente da faixa etária, género e região de residência, o consumo de peixe pela população

inquirida (população francesa com mais de 18 anos que consome peixe pelo menos duas vezes por semana) era equivalente às recomendações do governo francês.

Portugal é o país da União Europeia que mais consome pescado *per capita*, pelo que compreender os fatores que influenciam esta escolha pelos portugueses torna-se determinante para a formulação de modelos de previsão de procura. Neste seguimento, Cardoso and Nunes (2013) realizaram um inquérito aos padrões de consumo da população portuguesa, incidindo nas preferências por espécie de peixe, na frequência de consumo, na porção média consumida e nos métodos de confeção do produto. Concluíram que a preferência dos portugueses incide no peixe de origem selvagem em detrimento do proveniente de aquacultura. Em relação às condições de armazenamento, a refrigeração é preferível a todos os outros métodos. Também se verificam variações quanto às preferências de acordo com a região do país: as populações costeiras parecem preferir peixes selvagens, na zona sul (Algarve) consome-se mais sardinha e linguado, enquanto no arquipélago da Madeira a preferência reside sobretudo no peixe-espada preto. O estudo também realça que as preferências de peixe dos consumidores não coincide com o seu consumo, o que se poderá dever ao baixo preço inerente a algumas espécies de peixe e a alguns métodos de armazenamento, destacando o peixe em conserva.

Mais recentemente, Almeida and Vaz (2015) descrevem os fatores impulsionadores do consumo de peixe em Portugal e as suas consequências. Uma grande diversidade de espécies e de modos de preparação são duas das características deste consumo. O estudo conclui que existem cinco grandes fatores impulsionadores: a geografia do país, os recursos marinhos, a pesca, fatores sociais (religião e hábitos sociais) e políticos e três grandes tipos de consequências: ambientais (relacionadas com a sobrepesca), económicas e para a saúde.

2.2 Previsão de Vendas no Setor do Retalho

Nos últimos anos, a indústria do retalho enfrenta dois grandes desafios: o crescimento das compras *online* e o aumento da consciência ambiental, provocando assim constantes alterações neste mercado. Com o objetivo de enfrentar essas mudanças, Fildes and Kolassa (2019) analisam os sistemas de suporte à previsão de procura e a sua importância no mundo do retalho. Enfatizam que as decisões do ponto de vista estratégico, tático e operacional numa empresa muitas vezes são suportadas pela previsão da procura. Por esse motivo, as empresas cada vez mais investem no desenvolvimento e melhoria dos sistemas de previsão, com o objetivo de ultrapassar o método anterior: o uso da intuição na hora da decisão. O estudo levanta algumas questões metodológicas no processo: a "necessidade de métodos adequados à procura intermitente", o "crescente uso de métodos de *machine learning* mas sem validação científica" e o "uso do julgamento pessoal em muitas decisões". No futuro, supor-se-á fundamental o uso de ferramentas de ML em modelos de previsão da procura na indústria do retalho. Contudo, o estudo enfatiza duas questões importantes sobre estas ferramentas: o "entendimento por parte dos utilizadores dos pormenores incluídos na implementação de modelos de ML" e a "necessidade de adicionar valor humano no processo".

De forma a compreender as causas de desperdício alimentar, Mena and Yurt (2011) mostram no seu estudo as conclusões de entrevistas realizadas a 43 retalhistas da indústria alimentar de dois países, Espanha e Reino Unido. Uma das principais causas identificadas pelos autores é o erro nos modelos de previsão de vendas, muitas vezes associado a fatores externos. Um modelo de previsão de vendas preciso é fundamental para uma gestão eficaz da cadeia de abastecimento, eliminando a incerteza e criando estabilidade, sendo crítico a sua implementação e desenvolvimento para qualquer organização. Desta forma, vários estudos têm sido realizados de forma a verificar quais as técnicas mais eficazes.

Com o crescente aparecimento dos métodos de ML em projetos de previsão de vendas cria-se uma dicotomia entre os métodos mais tradicionais e estes mais recentes. Alon and Sadowski (2001) comparam uma rede neural artificial (ANN) com métodos mais tradicionais *Winters Exponential Smoothing*, *Box-Jenkins* ARIMA e Regressão Multivariada para prever as vendas agregadas no contexto do retalho norte americano. O artigo apresenta duas grandes vantagens do uso de ANN em relação a outros métodos: a "capacidade de modelar relações mais complexas, até mesmo casos sem linearidade" e a "habilidade de superar as limitações dos modelos tradicionais no que diz respeito a *outliers*, erros de especificação e suposição de linearidade". No teste realizado no estudo, o desempenho do modelo ANN é melhor comparativamente aos outros. Dos métodos estatísticos implementados, o ARIMA apresenta melhores resultados. O modelo realizado no estudo comprova também que o método ANN consegue capturar a tendência dinâmica não linear, os padrões sazonais e as suas interações.

Diversos estudos depreendem que o desempenho dos modelos ANN na previsão da procura futura é superior em detrimento a métodos mais tradicionais. Para além do estudo apresentado anteriormente, também Bhadouria and Jayant (2017) provam essa conclusão. Com a utilização de um caso de estudo real de uma empresa de fabricação de engrenagens pretende-se avaliar o desempenho dos métodos de ML. Mais uma vez, o método ANN apresenta melhores resultados em comparação a outros tradicionalmente usados para modelar séries temporais, como *Exponential Smoothing*.

Diversas técnicas e modelos têm sido aplicados para a previsão de vendas e procura no setor do retalho, onde a importância destes é enorme, já que são a base do sistema de reabastecimento de cada produto. Nos últimos anos, têm sido desenvolvidos sistemas inovadores, como é o caso de sistemas híbridos que combinam diferentes técnicas e as vantagens de cada uma delas. O artigo de Aburto and Weber (2007) apresenta um sistema híbrido inteligente que combina dois métodos, o ANN, que como referido apresenta um bom desempenho, e o ARIMA. O objetivo do modelo híbrido é juntar os pontos positivos dos dois métodos. Cria-se assim um modelo de previsão de procura que é aplicado a um supermercado chileno. Os resultados revelam que o sistema híbrido apresenta melhor precisão do que os dois em separado.

Procura Censurada

Amjad and Shah (2017) pretendem estimar a procura efetiva numa empresa que vende guarda-chuvas e para tal relatam alguns problemas que os dados podem apresentar. As observações das

vendas podem ser ruidosas, incompletas ou censuradas, de forma que a procura efetiva pode ser difícil de estimar. O ruído deve-se a erros aleatórios na medição ou manutenção de registos. Dados incompletos devem-se a relatórios perdidos ou falhas a reportar valores. No que respeita ao facto dos dados serem censurados, isto acontece dado não haver informação das pessoas que poderiam querer comprar, porém não conseguiram devido a *stock-outs*. A procura efetiva (sem censura) vai variar de semana para semana, tornando-se ainda mais difícil de calcular. Ao ignorar os efeitos da censura na estimativa da procura, a procura resultante vai ser tendenciosamente menor que o valor verdadeiro. Como as previsões de vendas, habitualmente, são realizadas sem o valor da censura este problema pode impactar as vendas futuras. O papel da inteligência artificial pode ser preponderante para auxiliar a ultrapassar estes impactos, promovendo um planeamento de inventário mais preciso.

Não só a área do retalho pode sofrer de procura censurada. Albiński and Minner (2018) abordam um sistema de partilha de bicicletas híbridas, onde o desempenho é medido pelo número total de alugueres. Segundo os autores, "a literatura negligencia a procura censurada", portanto estes tentam encontrar uma forma de a resolver. Uma estação de bicicletas pode estar sem *stock*, e consequentemente a procura efetiva pode ser superior à procura registada e os modelos de previsão podem subestimar a procura real. Assim, este estudo visa criar um método que estime as vendas perdidas e prevê a procura quando ocorrem *stock-outs*, para isso os padrões de vendas de períodos não censurados são extrapolados para os períodos de censura. Neste caso, eles utilizam a média correspondente ao mesmo dia, hora e quando o inventário de bicicletas não chega a zero. Esta análise é realizada nos pontos de paragem de bicicleta numa área semelhante.

Por sua vez, Ozhegov and Teterina (2019) realizam uma análise às vendas de massa numa loja regional russa. Neste contexto, os autores aplicam a Regressão Linear bem como as *Random Forests* a dois conjuntos de dados diferentes. Num deles consideram só os dias em que não houve procura censurada, ou seja, em que não ocorreu *stock-out* e o outro contém todos os valores com ou sem procura censurada. Os resultados revelam que os modelos em que existe consideração dos valores censurados das vendas apresentam melhor capacidade de previsão.

2.3 Previsão de Vendas

Previsão, de acordo com Hyndman and Athanasopoulos (2018), é a capacidade de antever o futuro com a maior precisão possível, considerando todas as informações disponíveis, incluindo os dados históricos e o reconhecimento dos eventos futuros que possam influenciar as previsões. Um bom modelo de previsão captura as mudanças que estão a suceder e assume que a maneira como estas mudanças ocorrem continuará no futuro. Uma organização necessita de desenvolver um sistema de previsão, identificando os problemas, testando métodos diferentes e selecionando o mais adequado para a situação em específico. Esta capacidade de treinar, testar e validar pode garantir o sucesso da previsão e da própria organização.

No que respeita à previsão de vendas, é importante entender quais são as causas que impactam as mesmas. Makridakis and Hyndman (1997) referem que existem duas categorias de fatores: os

externos, que não conseguem ser controlados, mais relacionados com o comportamento do consumidor e os internos, como promoções e campanhas de *marketing*, fatores que já são controlados pelas empresas. O impacto destes será sempre diferente em cada categoria de produto, cliente e loja. Para os consumidores, os fatores impactantes podem ser vários, como é o caso das condições atmosféricas, festividades, eventos desportivos, preço do produto e promoções. O grande desafio para as empresas na construção dos seus modelos de previsão de vendas é a capacidade de antever o comportamento dos consumidores face a estas variações.

Seguindo a linha de raciocínio de Hyndman and Athanasopoulos (2018) existem dois tipos de modelos de previsão de vendas: modelos causais e os *time-series*. Os modelos causais pressupõem que a variável a ser prevista, neste caso o volume de vendas de peixe fresco, pode ser explicada por uma relação causal com um conjunto de outras variáveis. Já os modelos *time-series* não utilizam variáveis exploratórias, mas sim determinam o padrão existente nos dados anteriores, neste caso nas vendas anteriores de peixe, e extrapolam para o futuro.

Tanto os modelos causais como os *time-series* apresentam vantagens e desvantagens. A sua escolha depende da situação em que são aplicados. De acordo com Makridakis and Hyndman (1997), quando o objetivo é uma aplicação mais facilitada de um método de previsão de vendas, deve ser utilizado um modelo de série temporal. Enquanto que se o objetivo é uma maior precisão, os modelos causais apresentam um melhor resultado na hora da decisão.

Boone and Sanders (2019) destacam o impacto tecnológico nas previsões e de que forma permite a melhoria. O potencial da quantidade de dados que agora pode ser recolhido e analisado quase em tempo real é enorme. Isto proporciona a melhoria do entendimento do comportamento do consumidor e desta forma impacta a precisão dos modelos de previsão de vendas. No entanto, o uso de tantos dados traz desafios inerentes, o tamanho e a desorganização dos mesmos pode ser desencorajador e são necessárias várias metodologias de previsão para combater a dispersão e não repetição dos dados. Ao nível de gestão, as empresas podem apresentar três contrariedades relacionadas entre si: a segurança, a privacidade e a regulamentação. Apesar destes desafios, o uso dos dados e da tecnologia revoluciona o mundo das previsões de vendas e os ganhos são enormes. O artigo perspetiva o futuro das previsões, onde as empresas podem usar grandes quantidades de dados em tempo real, conseguem ter uma imagem clara do comportamento do consumidor e uma previsão tão precisa que não melhora só a eficácia da cadeia de abastecimento mas também as receitas da instituição.

Capítulo 3

Caracterização das Vendas de Peixe Fresco

O presente capítulo apresenta as características principais da empresa retalhista que, ao fornecer um grande volume de dados das suas operações, possibilitou este estudo.

Na Secção 3.1 é exposta uma introdução à empresa parceira bem como a descrição do seu setor de pescado. Esta secção é seguida da caracterização dos dados na Secção 3.2, com uma análise da sazonalidade das vendas.

3.1 Descrição da Empresa

A empresa parceira do estudo apresenta um crescimento contínuo no mercado do retalho alimentar português nos últimos 30 anos, encontrando-se diversas vezes na liderança da escolha dos consumidores. O início da atividade da unidade alimentar remete ao ano de 1985 com a criação do primeiro hipermercado em Portugal, em conjunto com uma empresa francesa. O grupo apresenta um vasto leque de soluções para os consumidores, com mais de 1200 lojas na Península Ibérica, com um portefólio multifacetado, desde hipermercados urbanos até supermercados de proximidade. A estratégia da empresa firma-se num conjunto de pontos diferenciadores: a oferta de preços competitivos; um compromisso de qualidade e variedade nos frescos; um programa de fidelização direcionado para as famílias portuguesas; a sua própria marca, com produtos exclusivos e uma aposta na vertente do *e-commerce*.

Uma das estratégias principais adotadas pela empresa é o cartão de fidelização, com a disponibilização de benefícios e promoções numa rede alargada de parceiros, e que conta já com mais de 3.5 milhões de famílias portuguesas aderentes. 85% das suas vendas são de clientes com este cartão. Assim, cria-se um elo de ligação à marca, potenciando as vendas e a confiança dos consumidores.

A empresa de retalho alimentar apresenta diversas áreas de produtos. O foco do estudo será as 107 espécies de peixe fresco, dentro destas 89 são de origem selvagem e 18 provenientes de aquacultura. Visto que a aquacultura é um processo industrializado, torna possível a existência

desta cadeia de abastecimento durante todo o ano. Pelo contrário, e de forma a preservar a vida selvagem, as restantes 89 espécies apresentam sazonalidade associada.

A empresa não faz previsão de vendas para a segunda-feira visto que não há abastecimento nesse dia da semana. Nesses dias também não são registadas vendas porquanto elas encontram-se considerados nos domingos anteriores. Portanto, na análise de dados não foram tidos em conta os dias correspondentes a uma segunda-feira.

Assim, e já que a frescura do produto é um ponto fundamental na hora de compra deve-se garantir que as ordens são feitas diariamente. No caso desta empresa de retalho, as previsões de vendas de peixe fresco são realizadas com dois dias de antecedência. Desta forma é possível prevenir que os produtos se estraguem e garantir que o elevado número de dados é devidamente tratado. A correta estimação da previsão de vendas é fundamental para prevenir *stock-outs*, evitar desperdícios e aumentar o lucro.

3.2 Caracterização dos Dados

Os dados referentes aos detalhes das vendas de peixe, promoções, preços e *stocks* foram fornecidos por uma empresa portuguesa de retalho alimentar. De forma a extrair o máximo de informação possível é fundamental entender as duas bases de dados disponibilizadas e as suas características.

Da primeira, cujos dados se encontram agregados por dia e loja de cada *SKU*, é possível extrair: quantidade real de venda regular das lojas; quantidade promocional de venda das lojas; quantidade prevista de venda do produto nas lojas e quantidade de *stocks* nas lojas. A quantidade vendida no total pela loja é possível obter somando as duas quantidades fornecidas, em formato regular e promocional. A segunda base de dados encontra-se agregada por semana para cada *SKU* e contém: preço real do produto; preço do produto com o desconto da semana e percentagem deste desconto.

O horizonte temporal das bases de dados fornecidas é de 2 anos e 10 meses, aproximadamente, desde 01/01/2017 até 22/10/2019. Estes dados respeitam às vendas de três lojas do país: a loja 3, em Alcabideche, Cascais, um hipermercado urbano; a loja 246, em Sintra, um supermercado de grande dimensão e a loja 326, na Areosa, Pedrouços, um supermercado de proximidade.

As bases de dados contêm milhares de observações, necessitando de um elevado nível de atenção e tempo para o seu estudo. De forma a realizar uma análise mais expedita e detalhada é fundamental reduzir o volume de dados. O Gráfico 3.1 mostra que a loja 3 apresenta um maior volume de vendas de peixe fresco, tal como era esperado, já que é um hipermercado de maiores dimensões, abrangendo mais famílias. Portanto, os dados usados no estudo serão sempre referentes a esta loja.

Com a aplicação de uma análise ABC, frequentemente usada na caracterização de inventário, em que este é dividido em 3 categorias A, B e C com valor decrescente, é possível desagregar os *SKU's* conforme o volume de vendas. Admitindo que os produtos A, B e C correspondem a, respetivamente, 70%, 90% e 100% do volume acumulado de vendas. Os produtos A correspondem

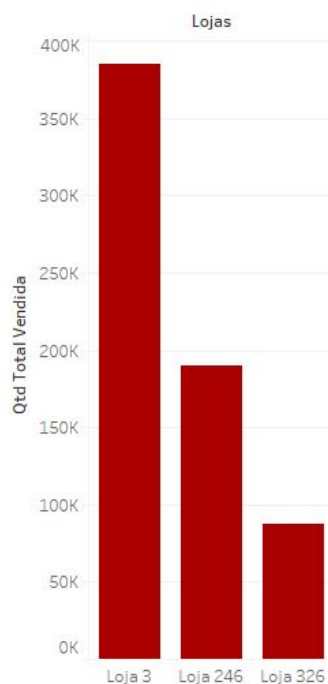


Figura 3.1: Vendas Totais de Peixe Fresco por Loja

a apenas 8% da variedade de *SKU's*, já os produtos B correspondem a 12% das espécies e os restantes 80% são da categoria C. No grupo A encontram-se peixes frequentemente mais conhecidos dos consumidores, como a dourada e o salmão, o que pode ser analisado no Gráfico 3.2.

Depois de analisar as escolhas dos consumidores na loja 3, hipermercado urbano em Cascais, compreende-se que a Dourada Média (200-600gr) é o peixe mais vendido. Por isso, será a espécie utilizada para os estudos iniciais.

O ponto de partida inicial para a análise da sazonalidade das vendas é uma visão mais macroscópica, demonstrada no Gráfico 3.3, onde se encontram presentes as vendas por trimestre do ano. Deste é possível retirar que as vendas são sempre superiores no terceiro trimestre, que corresponde aos meses de verão (julho, agosto e setembro). Este trimestre é caracterizado por uma temperatura superior, que potencia as refeições grelhadas ao ar livre, como é o caso do peixe fresco. Contudo, nos dois primeiros anos verifica-se que no segundo trimestre, os meses de primavera, esta tendência de crescimento já se encontra visível. Outro dado interessante analisar é o decréscimo de vendas em 2019 face aos dois anos anteriores. Tal pode acontecer devido a mudanças nas preferências dos consumidores, ou até o clima mais invernososo durante o ano (menor temperatura e mais precipitação). O último trimestre de 2019 encontra-se com valores tão baixos visto que não estão disponíveis os dados para todo esse período.

Já a um nível mais granular torna-se interessante analisar as vendas semanais da Dourada, no Gráfico 3.4. Nestes moldes é mais complicado verificar uma tendência de vendas, sendo no entanto notório que existem picos ao longo do ano e que estes geralmente se encontram associados a uma compra passada mais pequena. Estas variações podem se encontrar diretamente relacionadas com o preço dos produtos e as promoções semanais apresentadas pela empresa aos seus clientes.

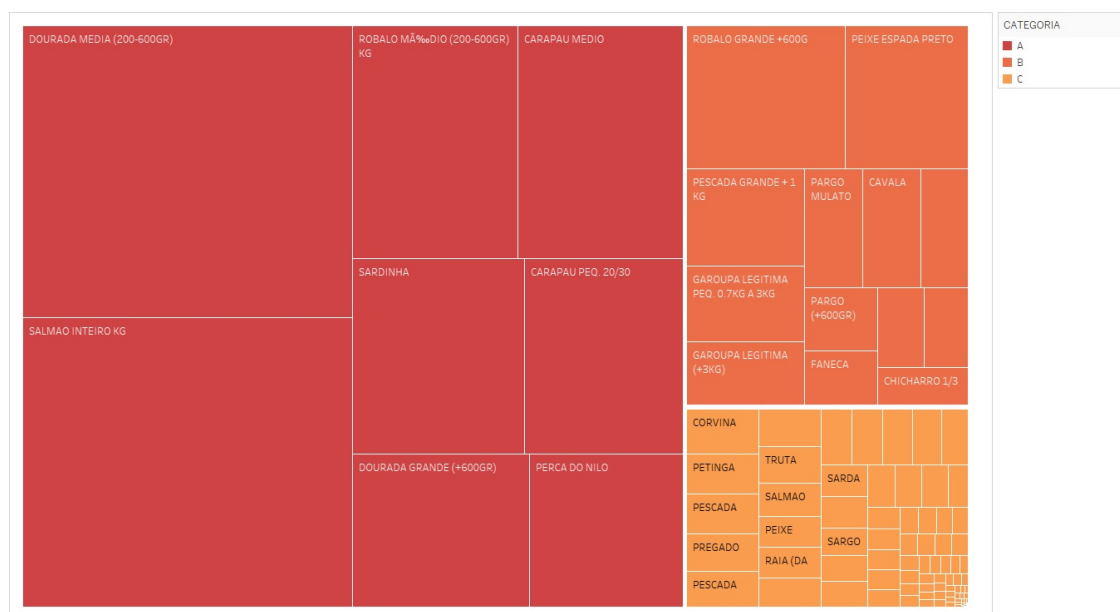


Figura 3.2: Treemap das Vendas de Peixe por SKU e categoria

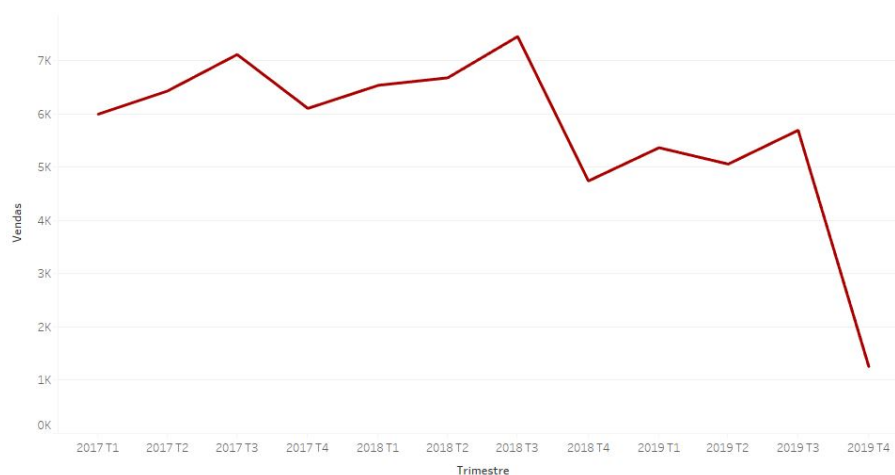


Figura 3.3: Vendas de Dourada Média por Trimestre

Pode, portanto existir uma relação forte entre a compra atual e a compra efetuada no mesmo dia da semana anterior, sendo uma questão necessária a ter em conta na realização dos modelos de previsão de vendas.

Outra questão importante e que se deve realçar é a sazonalidade ao longo da semana, representada no Gráfico 3.5. Tal como referido anteriormente, as vendas à segunda-feira são nulas visto que não são contabilizadas pela empresa e são contadas no domingo. Porém mesmo que estas fossem tidas em conta, era esperado não ser um valor muito elevado visto que os consumidores sabem que não há pesca de peixe selvagem ao Domingo, e como tal, este tipo de peixe não é fresco à segunda-feira. Outro aspeto que é interessante verificar é o facto do volume de vendas ser superior aos fins de semana e à sexta-feira, talvez porque é a altura que as pessoas apresentam uma maior disponibilidade para realizar compras. Esta constatação pode também advir do tipo de

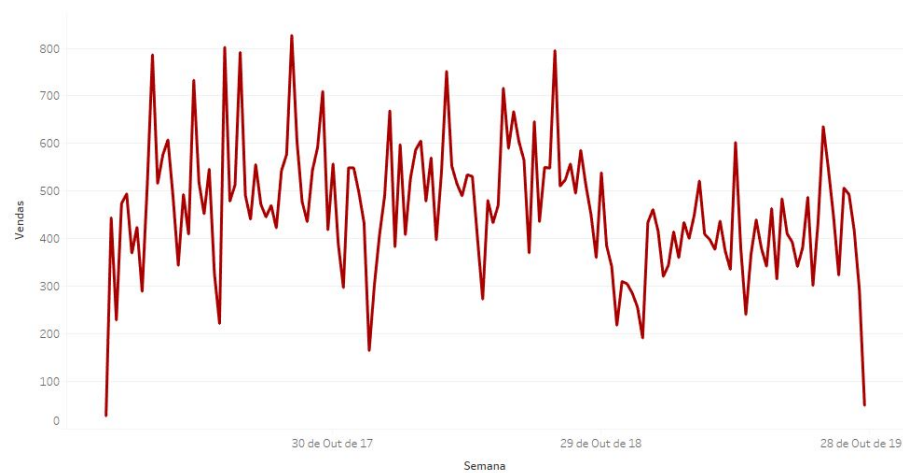


Figura 3.4: Vendas de Dourada Média por Semana

supermercado analisado, um hipermercado com uma área elevada, em que se realizam mais compras "semanais" familiares. Através desta análise é expectável que o dia da semana tenha também um impacto nos modelos, tal como a quantidade de peixe vendida no dia homólogo da semana anterior.

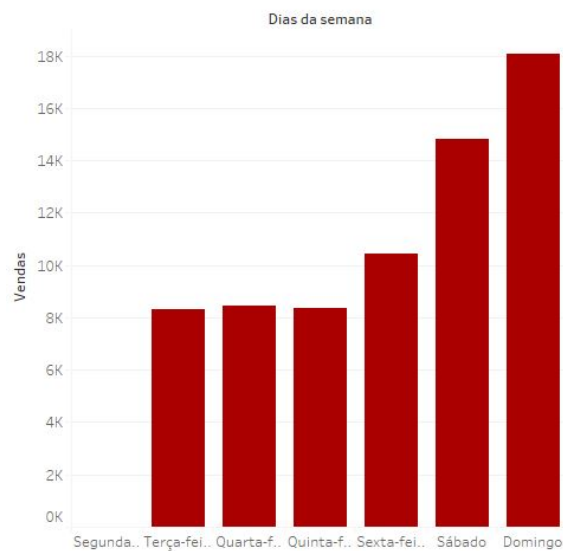


Figura 3.5: Vendas Acumuladas de Dourada Média por Dia

3.3 Conclusão

Tendo em conta as características do peixe, como é o caso da sua natureza perecível, e as especificações da cadeia de abastecimento, é importante uma previsão de vendas o mais precisa possível e realizada numa janela temporal estratégica. Neste caso, a empresa trabalha com dois dias de antecedência, sendo que só realiza estas previsões para seis dias da semana.

Os dados fornecidos referem-se a 2 anos e 10 meses, aproximadamente, para um conjunto de três lojas com características distintas. Na loja 3 realiza-se mais vendas em comparação com as outras duas, sendo consequentemente esta a loja objeto de estudo. A nível do volume de vendas das espécies é possível verificar uma divisão com um pequeno número de espécies mais vendidas, mas em claro contraste com as restantes. Destacando-se três principais, Dourada Média, Salmão Inteiro e Robalo Médio, sendo estas as utilizadas para a construção dos modelos.

Com base no padrão de vendas da dourada, e no que concerne à sazonalidade das vendas, assinala-se um padrão nas vendas mais elevadas, sempre correspondendo ao terceiro trimestre do ano. Já o ano de 2019 apresenta um volume menor de vendas face aos dois anos anteriores. Ao nível das vendas semanais visualizam-se picos de vendas ao longo das semanas, normalmente associados a valores mais baixos nas semanas anteriores. Já a um nível mais granular, compreende-se que as compras aos fins de semana são privilegiadas, e durante a semana tem um valor muito semelhante entre si, excetuando a sexta-feira. A partir destes dados é possível deduzir que tanto a quantidade vendida no dia homólogo da semana anterior como o dia da semana da compra terão influência no modelo.

Capítulo 4

Metodologia

O objetivo do projeto é o desenvolvimento de um modelo de previsão de vendas de peixe fresco para uma loja de retalho alimentar em Portugal. Este modelo visa ser o mais eficaz possível, com a capacidade de reduzir o desperdício do produto e também com a habilidade de evitar *stock-outs*. Sem esquecer que o primeiro grande intuito será arranjar uma heurística capaz de encontrar a procura real do produto, ou seja, aquela que considera a diferença entre vendas e procura e tem em consideração o nível de inventário diário.

Tal como se verificou no Capítulo 2 os modelos de previsão de vendas podem ser de duas categorias, causais e *time-series*. Neste projeto pretender-se-á comparar o desempenho de ambos, e como tal serão realizados três modelos de previsão de natureza causal e dois modelos de previsão baseados numa série temporal. Posteriormente, será realizada uma comparação dos resultados de cada modelo bem como o confronto com o modelo atualmente utilizado pela empresa. De tal maneira, procurar-se-á obter o modelo mais afinado possível e com um baixo erro associado.

No presente capítulo são explicadas todas as etapas necessárias para a realização do trabalho bem como o contexto teórico por detrás destas. Primeiramente, é executada a análise do processo de *Data Mining* depois disso cada etapa deste processo também será estudada, com foco nas decisões tomadas em cada uma.

4.1 Processo de *Data Mining*

Em primeiro lugar, é importante definir o que é *Data Mining*, segundo Han and Pei (2012): "processo de descobrir padrões interessantes e extrair conhecimento de uma grande quantidade de dados". As fontes destes dados podem ser "bases de dados, a Internet, armazéns de dados, outros repositórios de dados ou dados transmitidos diretamente para o sistema". Tendo em consideração a competitividade no mercado do retalho alimentar é fulcral o uso das ferramentas de *Data Mining*. Desta forma será possível analisar os dados da empresa possibilitando a perceção do perfil dos consumidores, das suas preferências e de que maneira os atrair.

Muitas vezes o termo de *Data Mining* encontra-se associado a todo o processo de *Knowledge Discovery in Databases*. Mas diversos autores consideram que este é apenas uma das etapas de todo o processo, como é possível visualizar na Figura 4.1.

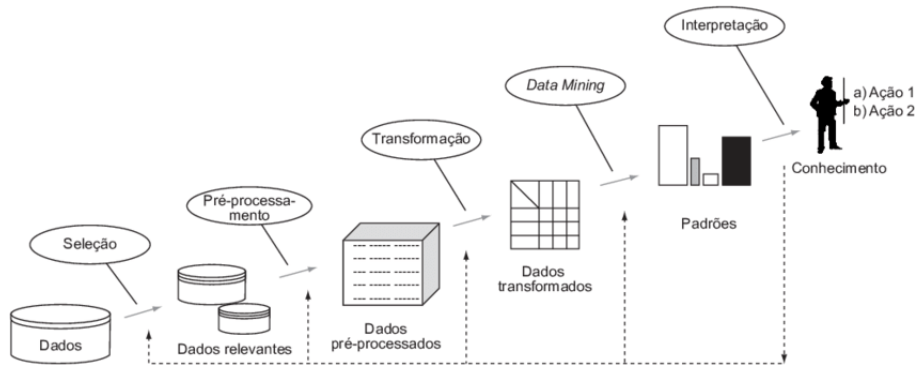


Figura 4.1: Visão geral dos processos de *Data Mining* (Steiner and Neto, 2006)

De acordo com Fayyad and Smyth (1996), o processo KDD apresenta as seguintes etapas:

- **Seleção:** Primeiramente é importante entender o domínio do problema bem como as metas a alcançar com o processo. Possibilitando a seleção dos dados importantes para este objetivo.
- **Pré-processamento:** Esta fase serve para trabalhar os dados de forma a serem usados na análise. Para isso, são feitas pequenas tarefas: remoção dos dados com ruído e inconsistentes, inferência dos dados ausentes, identificação e remoção dos *outliers*.
- **Transformação:** Este passo envolve a redução dos dados, encontrando os recursos necessários nestes, com a aplicação de métodos de transformação (agregação, generalização e normalização). Esta fase permite a seleção das variáveis que vão ser utilizadas no modelo.
- **Data Mining:** Este é um passo fundamental do processo e envolve a aplicação de métodos inteligentes para a extração de conclusões e padrões.
- **Interpretação:** Identificação dos padrões e resultados interessantes para a análise e avaliação destes.

No final deste processo é possível chegar ao conhecimento que foi proposto na etapa inicial e as técnicas de visualização e de representação auxiliam nesta tarefa. As ferramentas de *Data Mining* são fundamentais para as empresas, possibilitando um tratamento dos seus dados de forma a avaliar os seus clientes e atingir objetivos pré-definidos. A extração de padrões escondidos apresenta um grande potencial não só na hora da tomada de decisão mas também na definição dos dados necessários para o futuro. O uso das ferramentas de *Data Mining* possibilitou uma evolução enorme, permitindo que a tarefa dos analistas seja mais real e mais facilitada. Claro que o elevado volume de dados que as empresas possuem dificulta habitualmente o processo.

4.1.1 Seleção

Tal como referido no Capítulo 3 os dados utilizados no projeto foram providenciados por uma empresa de retalho alimentar português e são referentes às vendas de peixe fresco. Para o fim do projeto apenas serão utilizados os dados dos três peixes mais vendidos: Dourada Média, Salmão Inteiro e Robalo Médio. Isto deve-se ao facto da quantidade de SKU's disponíveis ser enorme e tendo em consideração o tempo computacional necessário para realizar e aperfeiçoar estes modelos. Os dados são referentes a apenas uma das lojas da empresa. O horizonte temporal da análise será os anos de 2017, 2018 e 2019 até ao dia 22 de outubro. Considera-se que os modelos desenvolvidos no âmbito do projeto serão facilmente extrapoláveis para as restantes espécies.

Como referido anteriormente serão utilizados modelos causais para a realização dos modelos de previsão de vendas, portanto será necessário adicionar variáveis explicativas. Isto é, variáveis que possam explicar o comportamento dos clientes e influenciem as vendas de peixe fresco. As variáveis relacionam as vendas com diversos fatores, nomeadamente:

1 - Fatores Temporais

- **Dia da semana:** No Capítulo 3 já se concluiu que o dia da semana em que é efetuada a compra pode ser uma variável preponderante para o modelo. Esta variável vai variar entre 1 e 7, sendo que não serão analisados os dias 2 que correspondem à segunda-feira.
- **Semana do ano:** Esta variável vai variar entre 1 e 53 e corresponde à semana em que é efetuada a compra. Permite acomodar as semanas de férias, tanto no verão como na altura de dezembro, e que podem impactar as vendas.
- **Mês:** Assumindo que há alturas do ano que mais potencializam o consumo de peixe, utiliza-se esta variável para verificar este comportamento, sendo que varia entre 1 e 12.
- **Primavera/verão:** Tal como verificado no capítulo anterior, os trimestres correspondentes a primavera e verão apresentam um pico de compras. O valor 1 corresponde a estas duas estações e o valor 0 às duas restantes.

2 - Fatores Meteorológicos

Fildes and Kolassa (2019) referem que as condições meteorológicas normalmente apresentam um grande impacto nas vendas, nomeadamente a precipitação e temperatura do dia de compra. Visto que os modelos são realizados com 2 dias de antecedência, os valores que os retalhistas podem utilizar são apenas previsões.

- **Precipitação:** Corresponde ao valor para o dia da compra e para o local da venda em mm.
- **Temperatura:** Esta variável é apresentada em graus *celsius* e será o valor médio diário.

3 - Festividades/Eventos

- **Quaresma:** Esta é uma altura extremamente associada à redução no consumo de carne e portanto pode potencializar o consumo de peixe. Varia entre 0 e 1, sendo 0 para altura não quaresma e 1 para a quaresma.

- **Feriado:** Muitos deles potenciam a ida das pessoas aos centros comerciais, nomeadamente a compra de produtos para a casa. Varia da mesma forma da anterior.

4 - Características do produto

- **Preço do produto com o desconto da semana:** Uma das variáveis que caracteriza o produto e que maior relação pode ter com a compra.
- **Promoção:** A percentagem de desconto, que varia de semana para semana, pode também influenciar a compra de peixe fresco.

5 - Campanhas Promocionais Anteriores

- **Quantidade vendida no dia homólogo da semana anterior:** Outra ilação retirada do capítulo anterior é que os picos de compra encontram-se associados a um valor baixo na semana anterior. Portanto, a quantidade vendida no dia homólogo da semana anterior pode ter uma influência no modelo.
- **Quantidade vendida 2 dias antes:** Visto que a previsão é feita com um horizonte temporal de dois dias, será interessante analisar a influência da quantidade vendida no momento da previsão.
- **Preço no dia homólogo da semana anterior:** Seguindo a linha de raciocínio da quantidade vendida, também o preço a que o peixe foi anteriormente vendido pode influenciar as vendas.
- **Preço 2 dias antes:** Tendo em conta o horizonte temporal da previsão, mas desta vez para o preço.
- **Promoção no dia homólogo da semana anterior:** Novamente na mesma linha de pensamento das vendas da semana anterior, mas desta vez para a avaliação da percentagem da promoção.
- **Promoção 2 dias antes:** Por último, a percentagem de desconto no momento da realização das previsões poderá também influenciar os modelos.

Para a construção dos modelos *time-series* e também dos modelos causais são necessárias as seguintes variáveis:

- **Quantidade vendida no total:** Soma da quantidade real e quantidade promocional de venda das lojas. Será utilizada com a transformação necessária para lidar com a procura censurada, como será explicado posteriormente.
- **Data de venda:** Apesar de não entrar diretamente no modelo será necessária para auxiliar na divisão dos dados em treino, validação e teste.

Por último, outras variáveis são utilizadas tanto para as heurísticas implementadas para lidar com a procura censurada como para as métricas de erro:

- **Quantidade prevista de venda:** A quantidade que a empresa de retalho perspetiva vender para cada dia para a espécie de produto naquela loja.
- **Quantidade de stock:** A quantidade de produto que existe na loja no final do dia.

A caracterização das variáveis utilizadas no projeto encontra-se no Anexo A. Uma nota importante referir nesta caracterização é a utilização das variáveis: dia da semana, semana do ano e mês como numéricas. Visto que estas variáveis podem ser utilizadas quer como numéricas, quer como categóricas, realizou-se um teste de comparação de desempenho com um tipo e outro.

4.1.2 Pré-processamento

Esta é uma etapa fundamental do processo e é essencial verificar que os dados ficam tratados de forma a não prejudicar o modelo. Neste caso em particular, não existiam dados em falta e em termos de *outliers*, isto é, valores atípicos, apenas algumas quantidades apresentaram valores negativos, portanto estes não foram considerados para a análise, uma vez que respeitam a devoluções (não sendo estas alvo de estudo neste projeto). Ao analisar a base de dados também se verificou que alguns dias não estavam presentes mas foi possível verificar que eram segundas-feiras, ou seja, dia para o qual não é efetuada previsão de vendas e cujas vendas do dia são consideradas no dia anterior. Por isso, estes valores em falta não geraram problema.

4.1.3 Transformação

No processo de transformação dos dados, o objetivo será lidar com a procura censurada e desta forma encontrar a procura real. Como explicado, a procura censurada considera que as vendas são iguais à procura, não tendo em conta os casos de falta de inventário. Assim os modelos de previsão da procura ao se basearem apenas nas vendas podem conduzir a uma subestimação da procura. Para isso não acontecer o presente trabalho propõe três heurísticas diferentes.

Primeiramente é importante referir os dois casos que podem conduzir a uma subestimação da procura:

- **Subestimação Total:** Nestes casos não existe *stock* do produto na loja e consequentemente não são registadas vendas. Este valor não especifica os clientes que se deslocaram à loja para comprar, mas não o fizeram pela falta de *stock*.
- **Subestimação Parcial:** Nestes dias, verificam-se vendas do produto na loja mas no final do dia o *stock* é nulo. Portanto, não considera os clientes que podem ter se deslocado à loja, no entanto já não encontraram o produto disponível.

Com o estudo realizado no Capítulo 2 foi possível verificar que os trabalhos existentes sobre a procura censurada são escassos. Por conseguinte, não há muito fundamento que permita construir as heurísticas, apesar disso o trabalho de Albiński and Minner (2018) apresenta *inputs* interessantes. Neste estudo, os autores utilizam as bicicletas numa área semelhante e realizam a média para o mesmo dia e hora. Visto que os dados do projeto são de três lojas com características muito

diferentes e em diferentes zonas do país, não é possível realizar esta análise. Em contrapartida, pode-se considerar dados na mesma loja, em períodos de tempo em que a procura é semelhante.

As heurísticas avaliadas para lidar com a procura censurada foram:

- **Heurística do Dia da Semana:** Nesta possibilidade, as vendas seguem uma distribuição semelhante conforme o dia da semana, como já foi equacionado anteriormente. Consequentemente, a procura real será a média das vendas daquela espécie, naquela loja para o dia homólogo da semana anterior e posterior.
- **Heurística do Dia de Semana Promocional:** Este caso é semelhante ao anterior, porém em vez de ir buscar automaticamente o valor das vendas dos dias homólogos, só o faz se for da mesma natureza, promocional ou não promocional. Explicando, se no dia em falta as vendas são promocionais, a heurística só faz a média com os dias homólogos em que as vendas também são promocionais, e vice-versa.
- **Heurística dos Dias:** Neste caso, a procura real é a média das vendas do dia anterior e seguinte ao dia em falta.

Nos casos em que a procura censurada se verifica no início ou fim do período de análise, não sendo possível realizar a média com o anterior e seguinte, consideram-se apenas um destes valores. Exemplificando, no último dia do horizonte temporal considerado no projeto, não será possível realizar a média entre o anterior e o posterior, pois este último não existe. Considera-se assim que o valor real da procura será só o anterior.

Nos casos de subestimação total da procura, o valor das vendas observado é substituído pelo valor da procura, inferido a partir das heurísticas. Já nos casos em que a subestimação é parcial, realiza-se uma comparação com o valor existente de vendas e se o valor inferido for superior ao existente procede-se à alteração, se não for mantém-se o valor da base de dados.

4.1.4 Técnicas de *Data-Mining*

As técnicas de *Data-Mining* são diversas e distintas umas das outras, dependendo do objetivo final da análise. De acordo com Tan and Karpatne (2018), estas técnicas podem-se dividir em dois grandes grupos: os métodos de previsão e os descritivos. Os métodos de previsão utilizam diversas variáveis na previsão de valores desconhecidos ou futuros de outra variável, denominada variável dependente. Em contrapartida, os métodos descritivos procuram descobrir padrões para melhor interpretar e relacionar os dados. Dentro destes dois grandes grupos existem diversas técnicas utilizadas e que têm objetivos diferentes, representadas na Figura 4.2.

Em relação aos modelos de previsão estes podem ser de classificação ou de regressão. Os primeiros são utilizados quando a variável dependente, ou seja, a variável objetivo é categórica discreta ao passo que se for contínua se devem utilizar os métodos de regressão. Ao nível da análise descritiva dos dados, quando o fim é agrupar os dados segundo grau de semelhança deve se utilizar o *clustering*. Quando o objetivo é encontrar padrões que acontecem em comum dentro

de um grupo de dados deve-se utilizar as regras de associação. Por último, a detecção de anomalias deve ser utilizada quando o propósito é encontrar observações raras e que são diferentes da maioria.



Figura 4.2: Técnicas de *Data Mining*

Os métodos de *Data-Mining* podem também ser classificados em aprendizagem supervisionada e não supervisionada. Na aprendizagem supervisionada, os criadores do modelo providenciam ao programa um conjunto de *inputs* e o *output* desejado. A função do modelo é relacionar as variáveis de forma a produzir o *output* desejado com o *input* dado. Os modelos de aprendizagem supervisionada utilizam técnicas de classificação e de regressão para prever uma dada variável. Enquanto que na aprendizagem não supervisionada apenas os *inputs* são conhecidos e são providenciados ao algoritmo, já o *output* é desconhecido. *Clustering* é um exemplo de um método de aprendizagem não supervisionada.

Tendo em conta os objetivos do projeto, os modelos utilizados são de previsão, mais concretamente de regressão, visto que o volume de vendas é uma variável numérica (contínua). Portanto, é considerado um método de *Data-Mining* de aprendizagem supervisionada. Seguidamente, serão descritos os modelos causais e *time-series* utilizados no projeto.

4.1.5 Algoritmos de *Data Mining*

4.1.5.1 Rede Neural Artificial

A rede neural artificial é um método de ML inspirado no funcionamento do cérebro animal. Este método é composto por diversos nós de entrada, o *input*, e um nó de saída, o *output*, ligados por conexões com um peso específico associado. Tal como os neurónios, os nós encontram-se ligados entre si e interagem através de sinapses. Estes sinais de entrada e saída com uma conexão com peso atribuído têm o nome de *perceptron* e encontra-se representado na Figura 4.3.

Os *inputs*, nomeadamente x_1 , x_2 , x_3 , podem tomar valores de 0 e 1. Através da soma ponderada de cada valor de entrada seguida da subtração do valor definido como tendência, o *perceptron* definirá o resultado de 1 ou -1. Desta forma, este método só permite a resolução de problemas de classificação, sendo inadequado para questões de regressão. No entanto, é possível alterar o método para ser viável um resultado com uma variável numérica. Isto é possível através da utilização de uma função de ativação diferente de uma função de sinal. De tal forma, é exequível modelar

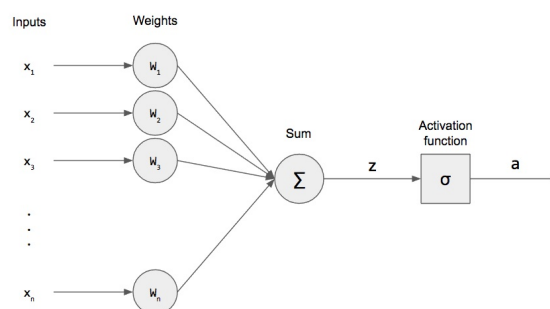


Figura 4.3: Funcionamento de um *Perceptron* (Deshpande, 2017)

uma variável discreta, possibilitando a utilização do método ANN em problemas de regressão, com o uso das seguintes funções de ativação: Lineares, Sigmóides ou Hiperbólicas.

Os sistemas neurais apresentados são adaptáveis e são baseados na experiência, utilizando processamentos não lineares. Estes modelos conseguem extrair padrões bastante complexos dos dados, visto combinarem um conjunto diverso de nós.

Na Figura 4.4 é possível observar o funcionamento apenas com uma camada embora seja possível utilizar este método com múltiplas camadas, utilizando os nós de saída de um *perceptron* como nós de entrada da próxima camada. Neste caso, os sistemas apresentam camadas escondidas que por sua vez têm um número arbitrário de nós e que servem de *input* à camada seguinte. Na prática, os modelos mais utilizados de ANN são os que apresentam uma camada escondida, como é possível observar o exemplo da Figura 4.4.

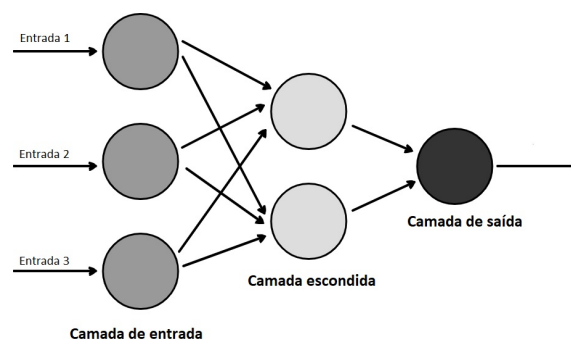


Figura 4.4: Representação de uma Rede Neural Artificial (Miller, 2020)

4.1.5.2 *Support Vector Regression*

O SVR é um método de ML com os mesmos princípios que o SVM, com apenas ligeiras diferenças, sendo utilizado nos problemas de regressão. No caso do SVM, quando confrontado com diversos dados de entrada que pertencem a uma de duas possibilidades num ambiente com n dimensões, constrói um subespaço com $n-1$ dimensões (hiperplano) que separa os dados em duas categorias. O modelo define o hiperplano utilizando vetores de suporte e margens, definidas pelos vetores. Os vetores de suporte são os dados de treino que se encontram mais próximos da margem

máxima do hiperplano. Este procura maximizar a distância entre os pontos mais próximos de cada uma das categorias. Esta maximização do hiperplano com a separação das duas categorias pode ser observada na Figura 4.5. Na escolha do hiperplano utiliza-se o que proporciona uma maior margem, visto que desta forma permite ter melhores erros de generalização, permitindo assim classificar os exemplos que ainda não foram observados.

Para ser possível tratar problemas de regressão existe uma adaptação do método descrito acima. Nas restrições para a construção do hiperplano, define-se que existem pontos em ambos os lados, mas a distância entre esses pontos e a linha não deve ser maior que a margem de tolerância. Desta forma, cria-se um hiperplano no meio de um conjunto de pontos, tornando-os o mais próximo possível. O foco do SVR é minimizar o erro de generalização para obter melhor desempenho, em vez de se concentrar em minimizar o erro de treino através da procura do hiperplano de margem máxima.

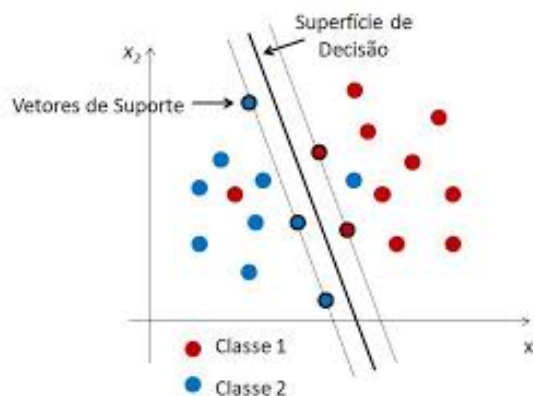


Figura 4.5: Representação de um hiperplano criado pelo método SVM (Silva and Santiago, 2017)

4.1.5.3 *Random Forest*

Random Forest corresponde a uma combinação de diversas árvores de decisão, sendo assim um método com diversos algoritmos desta forma consegue atingir um melhor resultado que apenas um algoritmo sozinho. Cada árvore é gerada baseando-se num conjunto dos dados originais, sendo cada conjunto independente entre si. O carácter aleatório ajuda a reduzir a correlação entre as árvores de decisão, o que significa que o erro de generalização deste método combinado consegue ser melhorado. O facto de este ser um método combinado de diversos algoritmos leva a uma grande precisão do resultado. Este processo pode ser observado na Figura 4.6.

Para compreender o funcionamento deste método é imprescindível analisar o funcionamento de uma árvore de decisão por si só. Uma árvore de decisão é composta por: um nó de raiz, que representa toda a população e que mais tarde é dividido em duas ou mais amostras; nós de decisão, quando um sub-nó é dividido em mais sub-nós e uma folha de cada árvore, quando os sub-nós não voltam a ser divididos. Este método abrange duas fases: a construção da árvore, que envolve a divisão de um nó em diversos sub-nós e a "poda" da árvore, que serve para remover ruídos e

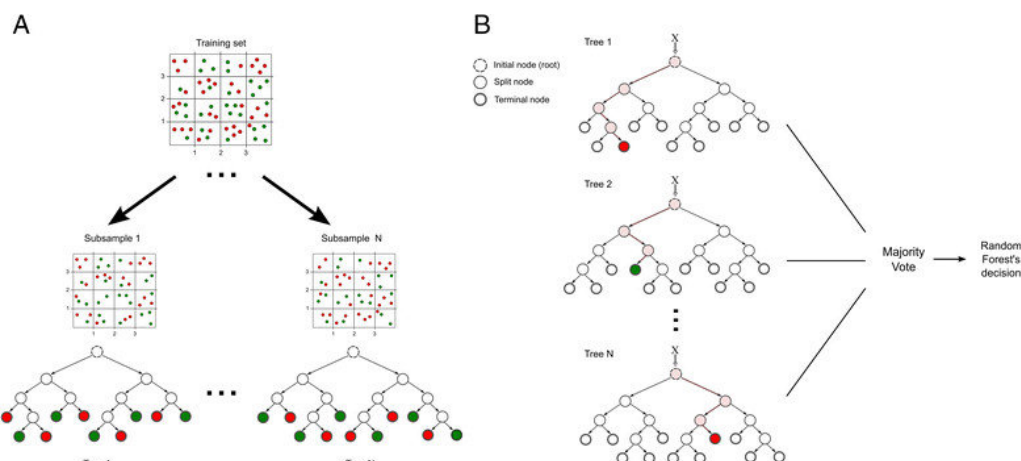


Figura 4.6: Representação de um modelo de *Random Forest* (Machado and Corbellini, 2015)

outliers. Para a utilização das árvores de decisão em problemas de regressão, o valor nas folhas das árvores é a média das observações dos dados colocados nessa folha.

4.1.5.4 SARIMA

SARIMA é um modelo muito semelhante ao ARIMA, Modelo Auto-regressivo Integrado de Médias Móveis, porém com a capacidade de lidar com a sazonalidade. Os modelos ARIMA são modelos estatísticos muito utilizados para a previsão de vendas. O ARIMA resulta da combinação de um modelo auto-regressivo e de um modelo de médias móveis. Nos modelos auto-regressivos, o valor atual da variável de interesse é calculado como um agregado finito e linear dos valores anteriores da variável. Por outro lado, os modelos de médias móveis não utilizam os valores passados da variável numa regressão, mas sim os erros de previsão anteriores num protótipo semelhante à regressão.

O modelo SARIMA é aplicado quando os dados manifestam um comportamento repetido ao longo do tempo, com similaridades depois de m períodos de tempo. O ARIMA Sazonal (SARIMA) é descrito da seguinte forma: $ARIMA(p,d,q) \times (P,D,Q)_m$, sendo que P,D,Q correspondem aos padrões sazonais de p,d,q , respectivamente. Por sua vez: p é a ordem da parte auto-regressiva, d é o grau da primeira diferenciação e q a ordem da parte de médias móveis. (Hyndman and Athanasopoulos, 2018)

4.1.5.5 Exponential Smoothing

Este método foi proposto no final dos anos 50, tendo a contribuição de três especialistas, Brown (1959), Holt (1957) e Winters (1960) e já motivou diversos métodos com sucesso. *Simple Exponential Smoothing* (SES) é o modelo mais simples deste método e o primeiro a ser criado. Este método realiza modelos de previsão de vendas que não apresentem nem tendência nem sazonalidade. Para ultrapassar estes obstáculos, Holt criou um modelo baseado no SES mas que

permitisse trabalhar os dados com tendência. Para isso, o autor criou uma equação de previsão com duas parcelas, uma para o nível e outra para a tendência.

Mais tarde, Holt e Winters juntaram-se para aperfeiçoar o método do primeiro, criando um modelo que lidasse não só com a tendência mas também com a sazonalidade. Deste trabalho em conjunto nasceu o modelo HoltWinters, que numa equação de previsão junta três: uma para o nível, l_t ; uma para a tendência, b_t e outra para a sazonalidade, s_t . Existem duas variações deste método que diferem na natureza do componente sazonal: método aditivo e método multiplicativo. O primeiro método é preferível quando as variações sazonais ocorrem de uma forma constante ao longo da série temporal. O método multiplicativo é utilizado quando as variações sazonais alteram proporcionalmente à série temporal. (Hyndman and Athanasopoulos, 2018)

4.1.6 Validação dos Modelos

4.1.6.1 Validação Cruzada

Para a implementação do modelo é necessário dividir os dados em pelo menos dois conjuntos, os dados de treino e os dados de teste. Os dados de treino são utilizados para ajudar a construir o modelo e os dados de teste são utilizados para a avaliação do modelo. Porém, pode-se acrescentar um terceiro conjunto de dados para a afinação dos parâmetros, ou seja, os dados de validação e que são independentes dos restantes.

No caso dos modelos causais, adotou-se uma abordagem *bootstrapping*, em que se construiu X amostras para efeitos de treino, validação e teste. O período que foi usado para criar as amostras de treino antecede o período usado para o período de validação e este antecede o período usado para constituir as amostras de teste. Este tipo de validação só acontece nos modelos causais.

4.1.6.2 Métricas de Erro

Para comparar os diferentes modelos realizados no projeto é necessário avaliar a sua precisão, comparando os resultados da aplicação dos modelos nos dados de teste com os valores reais de vendas. As medidas utilizadas na avaliação dos modelos encontram-se apresentadas em baixo, sendo baseadas em Adhikari and Agrawal (2013), e consideram $\hat{y}$ o valor previsto e y o valor observado.

- **Erro Absoluto Médio (MAE):** Mede o desvio absoluto médio entre os valores da previsão e os observados, sem considerar a direção dos erros. Este valor depende da escala de medição e da transformação dos dados. Uma boa previsão corresponde a um pequeno valor de MAE.

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (4.1)$$

- **Erro Percentual Absoluto Médio (MAPE):** Representa a média calculada dos erros percentuais absolutos pelos quais as previsões de um modelo diferem dos valores reais. Esta

medida é independente da escala de medição, contudo depende da transformação dos dados.

$$MAPE = \left(\frac{1}{n} \sum_{j=1}^n \frac{|y_j - \hat{y}_j|}{y_j} \right) \cdot 100 \quad (4.2)$$

- **Raiz quadrada do Erro Médio (RMSE):** Esta variável representa o desvio padrão dos resíduos (erros de previsão). Os resíduos são a medida que mede a distância dos pontos da linha de regressão.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (4.3)$$

4.1.7 Tecnologias Utilizadas no Projeto

Tendo em consideração o objetivo principal do projeto, o desenvolvimento de um modelo de previsão de vendas eficaz, será necessário utilizar tecnologias que permitam atingi-lo. Atualmente existem diversas tecnologias que permitem a implementação de métodos de previsão com aprendizagem supervisionada.

Na realização do projeto foram utilizados as seguintes tecnologias:

- **Tableau Desktop:** *Tableau Desktop* é um programa de visualização de dados que possibilita o entendimento dos dados e das suas relações rapidamente. A análise do problema executada no Capítulo 3 foi realizada com o auxílio deste *software* e foi utilizada a versão *Tableau 20.2*.
- **Python:** *Python* é uma linguagem de programação de alto nível e altamente funcional. Uma das suas principais características é permitir uma fácil leitura do código e exigir poucas linhas de código. Esta linguagem foi desenvolvida no programa *PyCharm Community Edition 2019.3.3* e foi utilizada para a transformação da variável da quantidade vendida de forma a lidar com a procura censurada.
- **R:** *R* é uma linguagem de programação, utilizada para a manipulação, análise e visualização dos dados. Esta linguagem é bastante usada por estatísticos e analistas aquando do desenvolvimento de um *software*. Esta linguagem foi utilizada para a realização dos modelos de previsão de vendas e para tal recorreu-se ao programa *RStudio*.

Capítulo 5

Resultados

Tendo presente a análise realizada no Capítulo 3 e seguindo a metodologia de trabalho descrita no Capítulo 4, neste será discutido a implementação das heurísticas e modelos do projeto. As heurísticas apresentadas encontram-se relacionadas com a procura censurada. Os modelos, estatísticos e de *machine learning*, procuram prever as vendas de peixe fresco, tendo por base a nova procura encontrada através da heurística selecionada numa primeira fase.

Na Secção 5.1 são apresentados os resultados das heurísticas implementadas nos valores censurados das vendas de peixe fresco. Já na Secção 5.2 são analisados os resultados dos cinco modelos implementados para cada espécie de peixe e a comparação com os valores da empresa de retalho. Na Secção 5.3 é avaliado o impacto da procura censurada nos modelos de previsão de vendas, com a comparação dos resultados quando a procura é e não é considerada censurada. Por último, na Secção 5.4 será estudado as variáveis que apresentam uma maior influência nas vendas.

5.1 Heurísticas para a Procura Censurada

Tal como explicado anteriormente, são considerados dois casos distintos de procura censurada, um em que a subestimação da procura é total e outro em que esta é parcial. Ao passo que na primeira os valores calculados pelas heurísticas substituem automaticamente o valor anterior, no caso da procura parcial só existe a substituição se o novo valor for superior ao anterior.

As heurísticas utilizadas são três e foram definidas tendo como base a análise realizada no Capítulo 3, visto que a literatura ainda apresenta poucas abordagens a este tema. Para a avaliação das heurísticas foram assumidos como desconhecidos 20% dos dados. Com a execução da heurística foram comparados os valores estimados com os valores iniciais.

Nas Tabelas 5.1, 5.2 e 5.3 encontram-se os resultados das heurísticas para as vendas censuradas de Dourada Média, Salmão Inteiro e Robalo Médio, respetivamente.

Tabela 5.1: Resultados das Heurísticas para as Vendas Censuradas de Dourada Média

Heurística	RMSE	MAE	MAPE
Dia da Semana Homólogo	15.28	4.08	13.60%
Dia da Semana Homólogo Promocional	16.11	4.36	15.10%
Dia Anterior e Seguinte	27.42	6.58	17.04%

Tabela 5.2: Resultados das Heurísticas para as Vendas Censuradas de Salmão Inteiro

Heurística	RMSE	MAE	MAPE
Dia da Semana Homólogo	16.06	4.44	10.27%
Dia da Semana Homólogo Promocional	16.21	4.47	10.39%
Dia Anterior e Seguinte	22.67	6.58	15.22%

Tabela 5.3: Resultados das Heurísticas para as Vendas Censuradas de Robalo Médio

Heurística	RMSE	MAE	MAPE
Dia da Semana Homólogo	6.30	1.74	8.88%
Dia da Semana Homólogo Promocional	6.46	1.80	9.20%
Dia Anterior e Seguinte	8.80	2.49	9.86%

Os resultados demonstram que, para as três espécies de peixe, a heurística do dia da semana homólogo, ou seja, a média da quantidade vendida no mesmo dia na semana anterior e seguinte, apresenta um menor valor de MAPE. Isto é, os erros percentuais absolutos entre os valores estimados e os aproximados são menores quando esta é a heurística aplicada. O mesmo acontece para as outras duas métricas utilizadas, MAE e RMSE.

Outra observação interessante prende-se com a diferença dos valores, principalmente nas métricas MAE e RMSE, para o caso do Robalo. Uma explicação plausível é o menor volume de vendas desta espécie de peixe. Tal como ficou visível, no Capítulo 3, a dourada e o salmão representam uma grande parte das vendas de peixe, desta loja, comparativamente ao robalo.

Com estas informações tidas em consideração, a métrica escolhida e utilizada para calcular a nova procura é a do dia da semana homólogo. Para a realização dos modelos de previsão de vendas mostrados de seguida, utiliza-se a procura efetiva calculada com recurso a esta heurística.

5.2 Modelos de Previsão de Vendas

Depois do pré-processamento dos dados e da transformação da variável quantidade vendida, descrita na secção anterior, o passo necessário é a implementação dos modelos de previsão de vendas.

Para a construção dos modelos foi necessário dividir os dados em treino, validação e teste, tendo em consideração os dados fornecidos pela empresa de retalho. Definiu-se que o ano de 2017 e os primeiros 6 meses de 2018 seriam os dados de treino, os restantes 6 meses de 2018 os dados de validação e o ano de 2019, até dia 22 de outubro, seriam os dados de teste.

5.2.1 Dourada Média

Com a implementação de dois métodos estatísticos e três métodos de ML, procedeu-se à comparação do seu resultado com o modelo de previsão utilizado atualmente pela empresa de retalho. Para cada um destes modelos e em cada uma das espécies, realizou-se a otimização dos parâmetros para permitir um resultado melhorado. Na Tabela 5.4 encontram-se os resultados dos modelos para a Dourada Média.

Tabela 5.4: Resultados dos Modelos de Previsão de Vendas de Dourada Média

Modelo	RMSE	MAE	MAPE
SARIMA	38.87	31.50	66.61%
HW	41.40	30.38	49.28%
ANN	32.86	23.96	42.91%
SVR	29.08	19.48	33.26%
RF	27.05	17.57	30.13%
Empresa de Retalho	29.57	20.62	33.65%

Para a implementação do método RF utilizou-se a função *randomForest()* presente no *R* através do pacote *randomForest*. Tal como explicado no Capítulo 4, este método gera diversas árvores de decisão, possibilitando a análise de diferentes formas do mesmo conjunto de dados. Neste método afinou-se dois parâmetros: *mtry*, número das variáveis aleatoriamente selecionadas por cada nó da árvore de decisão e *ntree*, número de árvores a crescer. Após o processo de afinação os valores adotados para o *mtry* foi 7 e o *ntree* 1500.

Já para a utilização do método *Support Vector Regression* utilizou-se a função *svm()* presente no pacote *e1071* do programa *R*. Para a implementação deste modelo utilizou-se o método radial e os parâmetros otimizados foram: *cost* e *gamma*. Para o melhor resultado do modelo, *cost* foi 2 e *gamma* 0.125.

Para a utilização do método ANN foi necessário utilizar a função *nnet()* presente no pacote do *R* com o mesmo nome. Este modelo baseia-se na construção de uma rede neural com diversas camadas, de entrada, escondidas e de saída, esta função utiliza uma camada escondida. Já ao nível dos parâmetros otimizaram-se: *size*, que define o número de unidades da camada escondida e *decay*, a capacidade do modelo se ajustar aos pesos da rede neural. Os melhores resultados conseguiram-se com um *size* igual a 60 e um *decay* de 0.0046.

Na implementação do método SARIMA foi necessário utilizar as funções *auto.arima()* e *Arima()* presentes no pacote *forecast* do *R*. O modelo aplicado e otimizado foi ARIMA(0,1,0)x(0,1,0)₆. Já para a implementação do método HW foi necessário utilizar a função *HoltWinters* presente no pacote *stats* do *R*. O parâmetro necessário para a aplicação deste método é *season* que descreve a sazonalidade dos dados. Para estes modelos utiliza-se o tipo *additive*. Nestes dois métodos estatísticos, os parâmetros não se alteram para cada espécie.

Considerando o MAPE, os modelos implementados com recurso às técnicas *Random Forest* e *Support Vector Regression* apresentam um melhor desempenho do que o modelo atualmente utilizado pela empresa de retalho. O método RF apresenta o melhor desempenho e a sua utilização

permite uma redução de cerca de 3% do erro percentual absoluto. Também os menores valores de RMSE e MAE são apresentados com a utilização deste método. Os dois modelos estatísticos utilizados, SARIMA e HW, apresentam pior resultado em comparação com métodos de ML.

5.2.2 Salmão Inteiro

Para esta espécie de peixe também foram realizadas iterações de forma a otimizar os modelos de ML implementados. Os resultados dos diversos modelos utilizados encontram-se na Tabela 5.5.

Tabela 5.5: Resultados dos Modelos de Previsão de Vendas de Salmão Inteiro

Modelo	RMSE	MAE	MAPE
SARIMA	30.08	22.70	34.50%
HW	35.98	25.89	37.58%
ANN	25.37	18.96	29.75%
SVR	24.49	17.51	25.58%
RF	19.80	13.98	20.80%
Empresa de Retalho	24.47	18.32	28.09%

Para a otimização do método RF utilizou-se um *mtry* igual a 11 e um *ntree* igual a 500. Já para a implementação do método SVR, o parâmetro *cost* utilizado foi 8 e o parâmetro *gamma* igual a 0.0020. Com um *size* de 10 e um *decay* de 1, o método ANN apresenta os seus resultados otimizados. Tal como explicado anteriormente, os parâmetros dos métodos estatísticos não foram alterados.

Mais uma vez, os modelos realizados com recurso aos métodos RF e SVR apresentam melhores resultados que o modelo atualmente utilizado. Desta vez, a técnica RF permite uma maior redução do erro percentual absoluto em comparação com o método atual, possibilitando um conjunto de vantagens substanciais para a empresa. Ao passo que na dourada, o método SARIMA apresentava o pior resultado e com uma discrepância substancial, o mesmo não acontece para esta espécie.

5.2.3 Robalo Médio

Por último, foram implementados os modelos de previsão às vendas de robalo médio e os resultados encontram-se na Tabela 5.6.

Na aplicação destes modelos otimizaram-se os diversos parâmetros de cada método. Para o RF utilizou-se um valor de 5 para o *mtry* e de 1000 para o *ntree*. Já para o SVR, o valor do *cost* foi 2 e o do *gamma* foi 0.0078. No método ANN, os valores dos parâmetros foram 10 para o *size* e 1 para o *decay*.

Tal como já tinha sido verificado, o método que apresenta melhor resultado ao nível de MAPE é o RF, seguido do SVR, e ambos apresentam um melhor desempenho que o modelo atualmente utilizado. Desta vez, o método de *Random Forest* permite uma redução de cerca de 5% do erro

percentual absoluto. Os métodos estatísticos apresentam novamente um pior resultado em comparação com os métodos de ML. Nos resultados desta espécie verifica-se um dado interessante, o reduzido valor de RMSE e MAE comparativamente às duas espécies anteriores. Tal como explicado nas heurísticas esta situação poderá acontecer devido ao menor valor do volume de vendas.

Tabela 5.6: Resultados dos Modelos de Previsão de Vendas de Robalo Médio

Modelo	RMSE	MAE	MAPE
SARIMA	17.59	13.39	47.60%
HW	19.89	15.17	49.39%
ANN	16.35	12.45	44.06%
SVR	14.94	10.80	35.58%
RF	13.00	9.68	33.60%
Empresa de Retalho	14.90	10.86	38.32%

5.3 Impacto da Procura Censurada

Para avaliar o impacto que a procura censurada, registada pela empresa nos modelos de previsão de vendas, procede-se à comparação dos valores previstos de vendas quando existe e não existe a consideração da procura censurada. Implementam-se os modelos da mesma forma descrita anteriormente, porém em vez de se utilizar a quantidade real, calculada na transformação dos dados, usa-se a quantidade observada pela empresa. Nesta comparação será utilizado o método com melhores resultados apresentados na secção anterior, o *Random Forest*. Nas Figuras 5.1, 5.2 e 5.3 encontram-se as comparações da quantidade prevista com a transformação da quantidade para lidar com a procura censurada e sem a transformação.

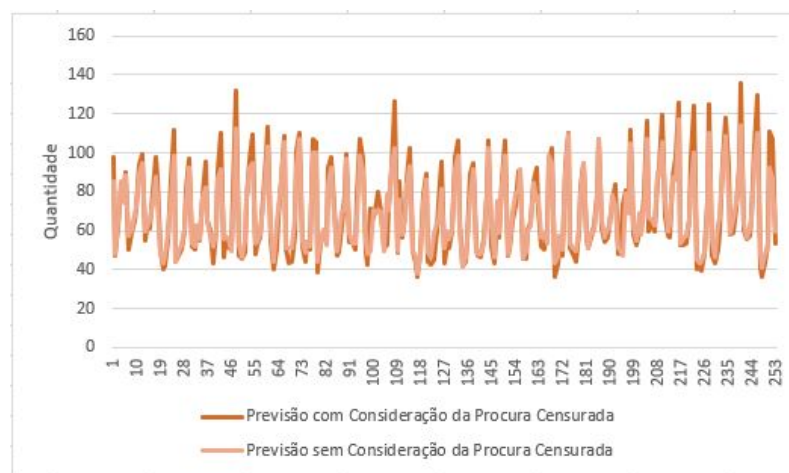


Figura 5.1: Comparação da Quantidade Prevista de Dourada Média Com e Sem o Tratamento da Procura Censurada

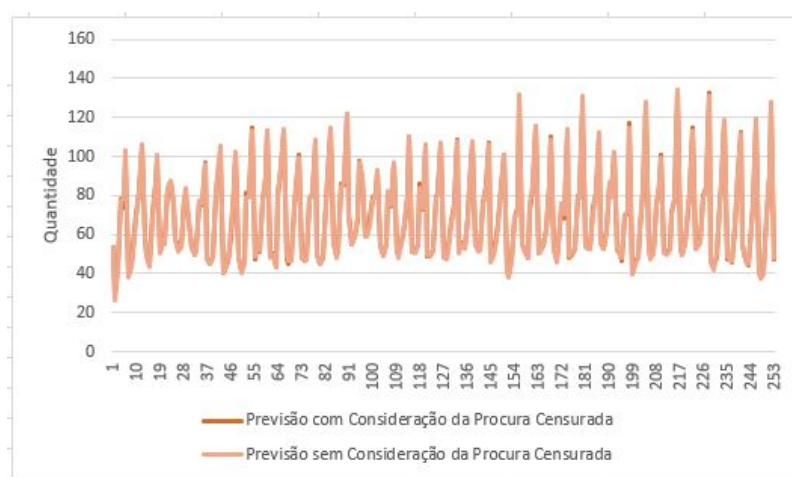


Figura 5.2: Comparação da Quantidade Prevista de Salmão Inteiro Com e Sem o Tratamento da Procura Censurada

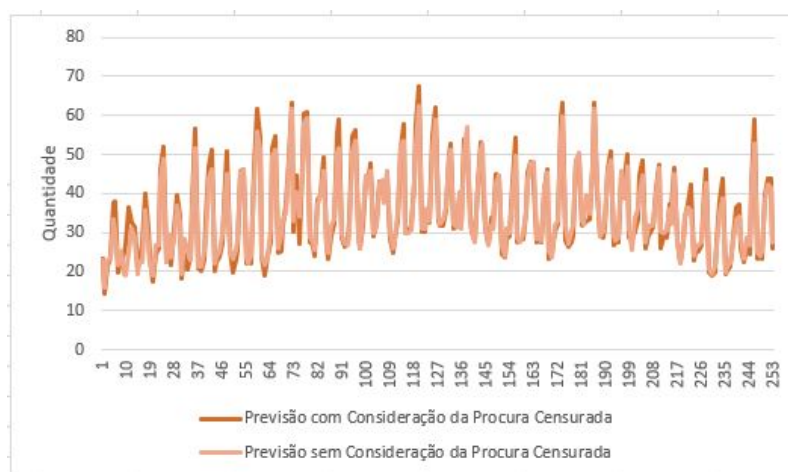


Figura 5.3: Comparação da Quantidade Prevista de Robalo Médio Com e Sem o Tratamento da Procura Censurada

Com a análise das Figuras, é interessante verificar que tanto para a Dourada Média como para o Robalo Médio, a quantidade prevista é superior quando ocorre correção da procura censurada. Em contrapartida, no caso do Salmão Inteiro, a quantidade prevista quando não é considerada a procura censurada é geralmente mais elevada do que quando é considerada.

De forma a verificar o desempenho dos modelos avalia-se cada modelo com recuso às métricas de erro. Nas Tabelas 5.7, 5.8 e 5.9 encontram-se os resultados da comparação dos modelos e ainda a comparação com o modelo atualmente utilizado.

Tabela 5.7: Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Dourada Média

Modelo	RMSE	MAE	MAPE
RF com consideração da procura censurada	27.05	17.57	30.13%
RF sem consideração da procura censurada	28.27	18.77	33.28%
Empresa de Retalho	29.57	20.62	33.65%

Tabela 5.8: Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Salmão Inteiro

Modelo	RMSE	MAE	MAPE
RF com consideração da procura censurada	19.80	13.98	20.80%
RF sem consideração da procura censurada	19.85	14.06	20.93%
Empresa de Retalho	24.47	18.32	28.09%

Tabela 5.9: Resultados da Comparação dos Modelos de Previsão de Vendas com o tratamento da Procura Censurada de Robalo Médio

Modelo	RMSE	MAE	MAPE
RF com consideração da procura censurada	13.00	9.68	33.60%
RF sem consideração da procura censurada	13.41	9.94	34.97%
Empresa de Retalho	14.90	10.86	38.32%

Um dado interessante analisar nestes resultados é o facto de os modelos mesmo quando não tratam a procura censurada apresentarem sempre melhor resultado que o atualmente utilizado. Consequentemente, a utilização deste método, RF, pela empresa de retalho já iria representar benefícios. No caso do Salmão, o ganho percentual na consideração da procura censurada e no seu tratamento é menor. Porém, a utilização dos dados originais sem transformação poderá resultar numa subestimação da procura, como pode ser verificado nas Figuras 5.1 e 5.3. De tal forma, a transformação das vendas em procura para lidar com os valores censurados é fundamental.

5.4 Importância das Variáveis

Uma das vantagens da utilização do método *Random Forest* é a capacidade do modelo de avaliar a importância das variáveis aquando da sua construção. No caso de problemas de regressão existem dois métodos que possibilitam esta avaliação de importância: um que se baseia no aumento percentual do Erro Quadrático Médio e é calculado baralhando os valores das amostras *out-of-bag* e outro baseado na diminuição da impureza quando uma variável é escolhida para dividir o nó da árvore. Neste projeto será utilizado o segundo método: a diminuição da impureza. Para problemas de classificação, utiliza-se a impureza do nó medida pelo índice de *Gini*. Em contrapartida, em problemas de regressão, a impureza é calculada pela Soma Residual dos Quadrados quando uma variável é escolhida para dividir o nó.

Na elaboração dos modelos em *R* é possível construir gráficos que mostrem a importância das variáveis com base na redução da impureza. As Imagens 5.4, 5.5, 5.6 são o resultado da análise.

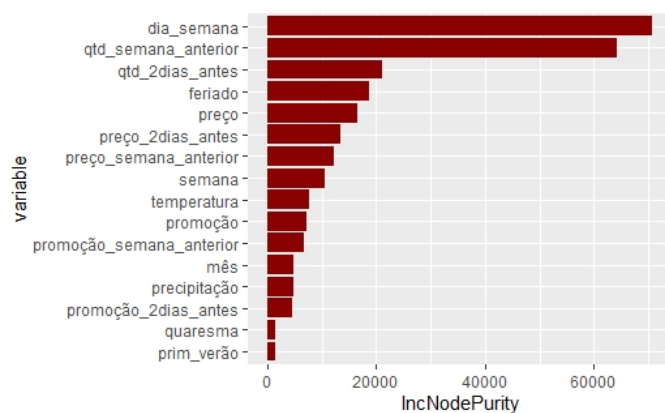


Figura 5.4: Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Dourada Média

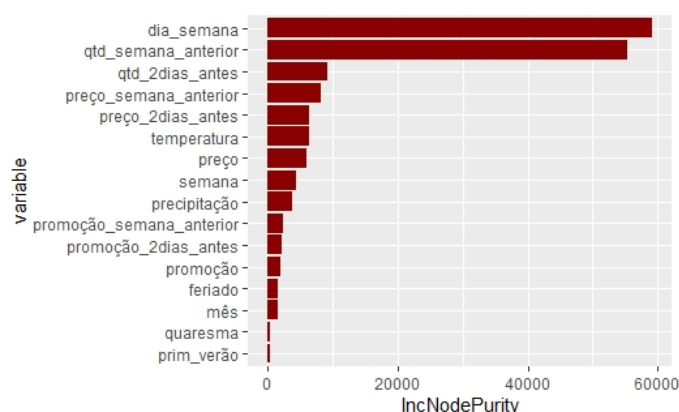


Figura 5.5: Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Salmão Inteiro

Nas duas primeiras espécies analisadas, o impacto das variáveis principais é superior comparativamente com o que acontece na espécie menos vendida das três, o robalo.

Uma conclusão comum aos três SKU's analisados é a importância de duas variáveis: dia da semana em que é efetuada a compra e quantidade vendida no dia homólogo da semana anterior. Já tinha sido anteriormente observado que o dia da semana poderia impactar as vendas. A utilização deste método permite perceber que é efetivamente uma característica com uma importância elevada. Estes resultados comprovam também o porquê da heurística dia da semana homólogo apresentar o melhor desempenho. Outra análise interessante é o fator feriado não apresentar uma grande importância na venda de salmão e robalo, mas ter uma grande importância na venda de dourada. No geral, o fator primavera-verão, a venda ser efetuada nestas estações do ano, é um dos menos importantes assim como o fator quaresma. Em contrapartida, a quantidade vendida dois dias antes, ou seja, no dia em que é realizada a previsão é relevante para as três espécies. Os três

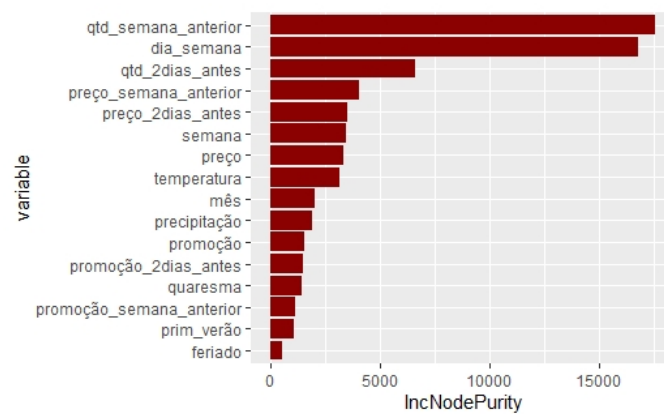


Figura 5.6: Importância das Variáveis na Construção do Modelo de Previsão de Vendas RF de Robalo Médio

fatores relacionados com o preço, preço atual, preço dois dias antes e preço do dia homólogo da semana anterior encontram-se nas variáveis mais importantes variando na sua ordem.

Nesta análise é possível verificar que apesar de uns fatores impactarem todas as espécies, as vendas variam muito de acordo com a espécie e os próprios fatores que impactam estas também não são sempre iguais.

Capítulo 6

Conclusão e Trabalho Futuro

Este estudo permitiu um conhecimento mais profundo do mercado de peixe fresco e uma comparação entre os diversos modelos de previsão de vendas. O objetivo primordial do projeto foi obter um melhor entendimento dos padrões de consumo dos clientes de pescado para a construção de um modelo de previsão de vendas eficaz e que seja capaz de lidar com a procura censurada.

Primeiramente e de forma a compreender a base de dados das vendas de peixe fresco fornecidas por uma empresa de retalho alimentar foi necessário realizar uma análise a todos os dados. A loja que apresenta um maior volume de vendas foi posteriormente a utilizada para a implementação do projeto. Para uma análise de todos os SKU's era necessário um grande esforço computacional, por isso foram escolhidas as três espécies com um maior volume de vendas na loja analisada: Dourada Média, Salmão Inteiro e Robalo Médio.

O primeiro grande propósito do projeto foi trabalhar a procura de forma a perceber quando é que ela é censurada e lidar com esses acontecimentos. Atualmente, os modelos de previsão de vendas utilizados pelas empresas e a própria literatura negligência estes casos, (Albiński and Minner, 2018), em que a quantidade vendida pode ser diferente da procura. Neste estudo foram identificados dois casos possíveis de subestimação da procura, um total e outro parcial. O primeiro ocorre quando não existem vendas daquela espécie, contudo também não existe *stock* e o segundo ocorre quando existem vendas, mas o *stock* esgota durante o dia. Estas duas possibilidades são tratadas de forma diferente, na primeira há uma substituição total do valor e na segunda só ocorre alteração quando o estimado é superior ao observado. Para lidar com estes problemas foram implementadas três heurísticas diferentes e testados os novos valores em comparação com os anteriores. A heurística que apresentou melhores resultados foi a do dia homólogo da semana, ou seja, a média entre a quantidade vendida no mesmo dia da semana anterior e seguinte. O novo valor de procura, calculado com recurso a esta heurística, foi o utilizado nos modelos de previsão de vendas implementados.

Numa fase seguinte, procedeu-se à implementação dos modelos de previsão de vendas e para isso selecionou-se três modelos de ML causais e dois modelos estatísticos *time-series*. Para o desenvolvimento dos modelos causais acrescentou-se variáveis consideradas relevantes na hora da compra de peixe fresco e que podem explicar o comportamento dos consumidores. Em cada

modelo de ML procedeu-se à otimização dos parâmetros para cada espécie.

Nas três espécies de peixe analisadas, o método *Random Forest* provou consistentemente permitir desenvolver o melhor modelo preditivo, com uma clara diminuição do erro percentual absoluto em relação ao modelo atualmente utilizado pela empresa de retalho. O método SVR também consegue apresentar melhores resultados que os atualmente implementados, no entanto, esta melhoria é menor que o caso do RF. A supremacia do RF encontra-se relacionado com a natureza do método, isto é, este resulta da combinação de diversos modelos (árvores de decisão), portanto apresenta melhor resultado que um único modelo. No que lhe concerne, os modelos estatísticos utilizados apresentam um desempenho menor em comparação com os modelos de ML, com um maior erro de previsão associado, tal como era expectável.

Num passo seguinte, foi fundamental comparar os resultados do modelo RF com e sem a transformação da procura, isto é, não considerar que existe procura censurada e utilizar diretamente a quantidade vendida nas lojas para a previsão das vendas. Esta etapa permitiu verificar que quando é considerada a procura censurada, a previsão de vendas é geralmente superior. Juntando os resultados da comparação dos desempenhos dos modelos através das métricas de erro torna-se evidente a importância de lidar com a procura censurada.

Por último, com a análise da importância das variáveis nos modelos de previsão de vendas para cada espécie é possível verificar que apesar de algumas serem notoriamente importantes para todos os SKU's outras variam de caso para caso. Por conseguinte, cada caso de venda de peixe é diferente dos restantes sendo importante uma análise granular e específica.

Seria interessante desenvolver a última análise como ponto de partida para a continuação deste projeto. Acredita-se que a utilização de um mecanismo de seleção das variáveis impactantes para cada espécie em específico e a utilização das mesmas nos modelos de previsão de vendas poderia vir a revelar-se promissora.

Para estudos futuros também seria interessante verificar quais os impactos das promoções nas preferências dos consumidores e de que forma existe canibalização neste mercado. Verificando também a existência de produtos substitutos para cada SKU e o impacto que estes podem ter na previsão do comportamento dos consumidores de peixe.

Por último e apesar de os modelos implementados neste projeto apresentarem bons resultados na generalidade, seria interessante avaliar se poderiam generalizar, isto é, quais os resultados para as restantes espécies, sobretudo as que apresentam um menor volume de vendas, e para as demais lojas.

Bibliografia

- Aburto, L. and Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1):136–144.
- Adhikari, R. and Agrawal, R. K. (2013). *An Introductory Study on Time series Modeling and Forecasting*. LAP LAMBERT Academic Publishing.
- Albiński, S., F.-P. and Minner, S. (2018). Performance analysis of a hybrid bike sharing system: A service-level-based approach under censored demand observations. *Transportation Research Part E*, 116:59–69.
- Almeida, C., K.-V. and Vaz, S. (2015). The seafood market in portugal: Driving forces and consequences. *Marine Policy*, 61:87–94.
- Alon, I., Q.-M. and Sadowski, R. (2001). Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8:147–156.
- Amjad, M. J. and Shah, D. (2017). Censored demand estimation in retail. *ACM SIGMETRICS Performance Evaluation Review*, 1(2).
- Bemrah, N., S.-V. L. J. and Volatier, J. (2008). Fish and seafood consumption and omega 3 intake in french coastal populations: Calipso survey. *Public Health Nutrition*, 12(5):599–608.
- Bhadouria, S. and Jayant, A. (2017). Development of ann models for demand forecasting. *American Journal of Engineering Research*, 6(12):142–147.
- Boone, T., G.-R. J. A. and Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, 35(1):170–180.
- Cardoso, C., L.-H. C. S. G. S. and Nunes, M. (2013). Survey into the seafood consumption preferences and patterns in the portuguese population. gender and regional variability. *Appetite*, 64:20–31.
- Claret, A., G.-L. A. E. R. L. H. M. M. I. P. J. G. A. and Rodríguez-Rodríguez, C. (2012). Consumer preferences for sea fish using conjoint analysis: Exploratory study of the importance of country of origin, obtaining method, storage conditions and purchasing price. *Food Quality and Preference*, 26(2):259–266.
- Deshpande, M. (09-2017). Perceptrons: The first neural networks.
- Expresso, J. (2017). Os portugueses são o povo europeu que come mais peixe. porquê?
- FAO (10-2017). Fishery and aquaculture country profiles, the portuguese republic.

- FAO (2018). *The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals*. 2018. FAO.
- Fayyad, U., P.-S. G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.
- Fildes, R., M.-S. and Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, pages 1–26.
- Han, J., K.-M. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 3rd edition.
- Hyndman, R. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, 2nd edition.
- IOM (2017). *Seafood Choices: Balancing Benefits and Risks*. The National Academies Press.
- Machado, G., M.-M. R. and Corbellini, L. G. (2015). What variables are important in predicting bovine viral diarrhea virus? a random forest approach. *Veterinary Research*, pages 46–85.
- Makridakis, S. G., W. S. C. and Hyndman, R. J. (1997). *Forecasting: Methods and Applications*. John Wiley Sons, 3rd edition.
- Mena, C., A.-D. B. and Yurt, O. (2011). The causes of food waste in the supplier–retailer interface: Evidences from the uk and spain. *Resources, Conservation and Recycling*, 6(55):648–658.
- Miller, S. (04-2020). Implementar redes neurais artificiais (rnas) no sql server.
- Ozhegov, E. M. and Teterina, D. (2019). Methods of machine learning for censored demand prediction. *Lecture Notes in Computer Science*, 11331(18):441–446.
- Silva, H. and Santiago, L. (2017). Aplicabilidade de técnicas de clusterização e máquinas de vetores de suporte para previsão de radiação solar em regiões tropicais.
- Steiner, M., S.-N. S. T. N. J. C. and Neto, P. J. (2006). Abordagem de um problema médico por meio do processo de kdd com ênfase à análise exploratória dos dados. *Gestão e Produção*, 13(2):325–337.
- Tan, P., S.-M. K. V. and Karpatne, A. (2018). *Introduction to Data Mining*. Pearson Education, 2nd edition.

Anexo A

Caraterização das Variáveis

Tabela A.1: Variáveis presentes nos modelos de Dourada Média

Atributo	Tipo	Descrição	Média(DP)/Valores(F)
1	Numérica	Quantidade Vendida no Total antes da Transformação	76.92(44.82)
2	Numérica	Quantidade Vendida no Total depois da Transformação	77.93(44.91)
3	Numérica	Quantidade Prevista de Venda	83.74(35.88)
4	Numérica	Quantidade de Stock	111.10(74.76)
5	Numérica	Dia da Semana	4.33(1.97)
6	Numérica	Semana do Ano	25.22(14.44)
7	Numérica	Mês	6.20(3.31)
8	Catégorica	Primavera/Verão	Não:45% Sim:55%
9	Numérica	Precipitação	0.82(2.61)
10	Numérica	Temperatura	17.97(4.76)
11	Catégorica	Quaresma	Não:86% Sim:14%
12	Catégorica	Feriado	Não:96% Sim:4%
13	Numérica	Preço do produto com o desconto da semana	5.35(0.27)
14	Numérica	Promoção	32.10(4.58)
15	Numérica	Quantidade vendida no dia homólogo da semana anterior	77.88(44.89)
16	Numérica	Quantidade vendida 2 dias antes	61.09(47.03)
17	Numérica	Preço no dia homólogo da semana anterior	5.35(0.27)
18	Numérica	Preço 2 dias antes	32.10(4.58)
19	Numérica	Promoção no dia homólogo da semana anterior	32.18(4.53)
20	Numérica	Promoção 2 dias antes	32.13(4.56)

Tabela A.2: Variáveis presentes nos modelos de Salmão Inteiro

Atributo	Tipo	Descrição	Média(DP)/Valores(F)
1	Numérica	Quantidade Vendida no Total antes da Transformação	76.24(40.99)
2	Numérica	Quantidade Vendida no Total depois da Transformação	76.78(41.15)
3	Numérica	Quantidade Prevista de Venda	80.41(33.27)
4	Numérica	Quantidade de Stock	179.37(95.16)
5	Numérica	Dia da Semana	4.33(1.97)
6	Numérica	Semana do Ano	25.22(14.44)
7	Numérica	Mês	6.20(3.31)
8	Categórica	Primavera/Verão	Não:45% Sim:55%
9	Numérica	Precipitação	0.82(2.61)
10	Numérica	Temperatura	17.97(4.76)
11	Categórica	Quaresma	Não:86% Sim:14%
12	Categórica	Feriado	Não:96% Sim:4%
13	Numérica	Preço do produto com o desconto da semana	8.68(0.46)
14	Numérica	Promoção	11.97(4.48)
15	Numérica	Quantidade vendida no dia homólogo da semana anterior	76.57(41.21)
16	Numérica	Quantidade vendida 2 dias antes	61.33(45.96)
17	Numérica	Preço no dia homólogo da semana anterior	8.69(0.46)
18	Numérica	Preço 2 dias antes	8.68(0.46)
19	Numérica	Promoção no dia homólogo da semana anterior	11.95(4.49)
20	Numérica	Promoção 2 dias antes	11.96(4.48)

Tabela A.3: Variáveis presentes nos modelos de Robalo Médio

Atributo	Tipo	Descrição	Média(DP)/Valores(F)
1	Numérica	Quantidade Vendida no Total antes da Transformação	30.90(18.19)
2	Numérica	Quantidade Vendida no Total depois da Transformação	31.20(18.14)
3	Numérica	Quantidade Prevista de Venda	38.34(17.42)
4	Numérica	Quantidade de Stock	78.13(51.59)
5	Numérica	Dia da Semana	4.33(1.97)
6	Numérica	Semana do Ano	25.22(14.44)
7	Numérica	Mês	6.20(3.31)
8	Categórica	Primavera/Verão	Não:45% Sim:55%
9	Numérica	Precipitação	0.82(2.61)
10	Numérica	Temperatura	17.97(4.76)
11	Categórica	Quaresma	Não:86% Sim:14%
12	Categórica	Feriado	Não:96% Sim:4%
13	Numérica	Preço do produto com o desconto da semana	5.51(0.37)
14	Numérica	Promoção	30.12(5.49)
15	Numérica	Quantidade vendida no dia homólogo da semana anterior	31.10(18.22)
16	Numérica	Quantidade vendida 2 dias antes	24.07(18.83)
17	Numérica	Preço no dia homólogo da semana anterior	5.51(0.37)
18	Numérica	Preço 2 dias antes	5.51(0.37)
19	Numérica	Promoção no dia homólogo da semana anterior	30.17(5.47)
20	Numérica	Promoção 2 dias antes	30.13(5.48)