

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO TECNOLÓGICO  
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS  
CURSO ENGENHARIA DE PRODUÇÃO MECÂNICA

Ian Vieira Silveira

**MODELO DE PREVISÃO DE DEMANDA COM O USO DE APRENDIZADO  
SUPERVISIONADO DE MÁQUINA: UM ESTUDO DE CASO EM UMA EMPRESA  
DE VAREJO**

Florianópolis

2019

Ian Vieira Silveira

**MODELO DE PREVISÃO DE DEMANDA COM O USO DE APRENDIZADO  
SUPERVISIONADO DE MÁQUINA: UM ESTUDO DE CASO EM UMA EMPRESA  
DE VAREJO**

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção Mecânica do Centro tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenharia, área Mecânica, habilitação Produção Mecânica.  
Orientador: Prof. Dr. Carlos Ernani Fries

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Silveira, Ian Vieira

Modelo de previsão de demanda com o uso de aprendizado  
supervisionado de máquina: um estudo de caso em uma  
empresa de varejo / Ian Vieira Silveira ; orientador,  
Carlos Ernani Fries, 2019.  
85 p.

Trabalho de Conclusão de Curso (graduação) -  
Universidade Federal de Santa Catarina, Centro Tecnológico,  
Graduação em Engenharia de Produção Mecânica, Florianópolis,  
2019.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. Engenharia de  
Produção Mecânica. 3. Aprendizado de máquina. 4. Análise de  
dados. 5. Previsão de demanda. I. Fries, Carlos Ernani.  
II. Universidade Federal de Santa Catarina. Graduação em  
Engenharia de Produção Mecânica. III. Título.

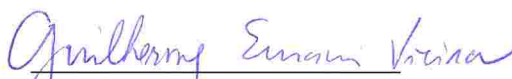


Ian Vieira Silveira

**MODELO DE PREVISÃO DE DEMANDA COM O USO DE APRENDIZADO  
SUPERVISIONADO DE MÁQUINA: UM ESTUDO DE CASO EM UMA EMPRESA  
DE VAREJO**

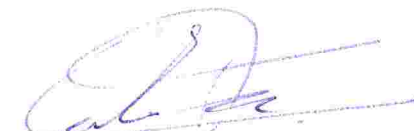
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Engenheiro Mecânico com habilitação em Produção e aprovado em sua forma final pelo Curso de Engenharia de Produção Mecânica

Florianópolis, 06 de dezembro de 2019.



Prof. Guilherme E. Vieira, Dr.  
Coordenador do Curso

**Banca Examinadora:**



Prof. Carlos Ernani Fries, Dr.

Orientador

Universidade Federal de Santa Catarina



Prof. Sérgio Fernando Mayerle, Dr.

Avaliador

Universidade Federal de Santa Catarina



Prof. Eduardo Ferreira da Silva Dr.

Avaliador

Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus amigos e aos meus queridos pais.

## **AGRADECIMENTOS**

Gostaria de agradecer a minha família, principalmente aos meus pais, Felipe e Eliane, e a minha irmã, Luana, que sempre me apoiaram e deram todas as condições para que eu me dedicasse aos estudos.

A meu orientador, prof. Carlos Ernani Fries, por ter me apresentado a área de análise de dados e por todas as horas de orientações durante esse período. Obrigado por todo o conhecimento compartilhado e por sempre me manter motivado a concluir este trabalho.

Aos meus amigos, por estarem junto comigo durante toda essa jornada, sempre me apoiando e torcendo por mim. Um agradecimento especial ao Alexandre Gomes e ao Gabriel Amante, que dividiram comigo muito dos momentos difíceis dessa graduação e também aos meus outros grandes amigos: José Brognoli, Unírio Machado, Ebert Rodrigues, Victor Diogo, Gabriel Volpato e Mirella Maia.

Ao Alexandre Daniel Scheidt e ao Carlos Henrique Tavares de Souza pela total disponibilidade quanto à solução de questões relativas à graduação e aos demais professores e servidores da UFSC pelo comprometimento e ensinamentos.

“Ciência é o conhecimento que nós entendemos tão bem que nós conseguimos ensinar ao computador. Todo o resto é arte.”

(Donald Knuth)



## RESUMO

Este trabalho tem como objetivo criar um modelo de previsão de demanda para itens de uma empresa de varejo. Um modelo que consiga prever a demanda futura com baixo erro de previsão permite que a empresa planeje melhor sua cadeia de suprimentos de forma a atender as necessidades dos clientes de disponibilidade de produtos e agilidade na entrega sem gerar um acúmulo elevado de estoques, o que diminui sua lucratividade. Nos últimos anos, os avanços na tecnologia e nos sistemas de coleta de dados resultaram na geração de grandes volumes de dados. A área de Análise de Dados tem ganhado grande destaque por ser capaz de extrair conhecimento desses dados e auxiliar organizações na tomada de decisões. Métodos de aprendizado de máquina são usados para identificar padrões em séries temporais e tem a capacidade de prever valores futuros com precisão cada vez maior. Para o desenvolvimento do modelo de previsão foram coletadas séries temporais com informações diárias de venda de quatro itens em todas as lojas da rede junto com informações de estoque, tipo de item, preço e estratégias promocionais. Com o uso do XGBoost, algoritmo derivado do método de florestas aleatórias, foi criado um modelo de previsão de demanda e os resultados obtidos foram comparados a modelos estatísticos (Holt-Winters e SARIMA), sendo que o modelo XGBoost apresentou menor margem de erros.

**Palavras-chave:** Previsão de demanda, análise de dados, séries temporais, aprendizado de máquina

## ABSTRACT

This work aims to create a demand forecast model for products from a retail company. A model that can predict future demand with low forecast error allows the company to better plan its supply chain in order to meet customer needs for product availability and agility in delivery without generating a high accumulation of inventory, which decreases its profitability. In recent years, advances in technology and data collection systems have resulted in the generation of large volumes of data. The area of Data Analysis has gained great prominence for being able to extract knowledge from this data and assist organizations in decision making processes. Machine learning methods are used to identify patterns in time series and have the ability to predict future values with increasing accuracy. For the development of the forecast model, time series were collected with daily sales information of 4 items in all stores of the chain along with information on stock, type of item, price and promotional strategies. With the use of XGBoost, an algorithm derived from the random forest method, a demand forecast model was created and the results obtained were compared to statistical models (Holt-Winters and SARIMA), and the XGBoost model presented a smaller margin of error.

**Keywords:** Sales forecasting, data analysis, time series, machine learning

## LISTA DE FIGURAS

Figura 1 - Trade-off custos x nível de estoque .....	22
Figura 2 – Resposta do consumidor à ruptura. ....	24
Figura 3 - Epíclio da análise de dados. ....	25
Figura 4 - Fluxograma dos tipos de questões em análise de dados.....	27
Figura 5 - Representação de uma árvore de decisão. ....	32
Figura 6 - Representação no espaço de uma árvore de decisão. ....	33
Figura 7 - Representação esquemática de <i>Bagging</i> em floresta aleatória. ....	35
Figura 8 - Esquema do <i>Boosting</i> . ....	37
Figura 9 – Interesse de usuários ao longo do tempo .....	39
Figura 10 – Exemplo árvore de decisão. ....	39
Figura 11 – Exemplo de <i>ensemble</i> em árvore de decisão.....	40
Figura 12 – Cálculo de score de estrutura .....	44
Figura 13 – Fórmula do ganho na divisão de folhas. ....	44
Figura 14 - Exemplo de divisão ideal.....	45
Figura 15 - Divisão do conjunto de dados em treino de teste.....	46
Figura 16 - Erros de previsão dos dados de teste e treino. ....	47
Figura 17 - Exemplos de <i>underfitting</i> , um modelo adequado e <i>overfitting</i> , respectivamente. ....	48
Figura 18 – Exemplo de validação cruzada usando <i>Backtesting</i> . ....	49
Figura 19 - Visualização da importância das variáveis com o SHAP.....	51
Figura 20 – Enquadramento da pesquisa. ....	52
Figura 21 - Fluxograma do roteiro do trabalho.....	53
Figura 22 - Localização das lojas e centros de distribuição. ....	56
Figura 23 - Itens selecionados para análise. ....	57
Figura 24 - Histórico mensal de vendas do item I (parafuso de fixação).....	57
Figura 25 - Histórico mensal de vendas do item II (tubo de ligação). ....	58
Figura 26 - Histórico mensal de vendas do item III (kit de instalação para aquecedor a gás).....	59
Figura 27 - Histórico mensal de vendas do item IV (assento sanitário).....	59
Figura 28 - Conjunto de dados inicial. ....	60
Figura 29 - Comando “info()” aplicado no conjunto de dados inicial. ....	61

Figura 30 - Dados estatísticos do conjunto de dados inicial.....	62
Figura 31 - Parte do conjunto de dados agrupado.....	62
Figura 32 - Participação dos itens dentro da sua família de itens. ....	63
Figura 33 - Decomposição do histórico mensal de vendas do item I.....	64
Figura 34 - Decomposição do histórico mensal de vendas do item II. ....	64
Figura 35 - Decomposição do histórico mensal de vendas do item III. ....	65
Figura 36 - Decomposição do histórico mensal de vendas do item IV.....	65
Figura 37 - Erros de teste e treino .....	71
Figura 38 – Importância das <i>features</i> no modelo de previsão. ....	73
Figura 39 Grau de importância das <i>features</i> com a ferramenta SHAP. ....	74
Figura 40 – Previsão do volume de vendas (em unidades) para o item I.....	75
Figura 41 – Previsão do volume de vendas (em unidades) para o item II.....	76
Figura 42 – Previsão do volume de vendas (em unidades) para o item III. ....	76
Figura 43 – Previsão do volume de vendas (em unidades) para o item IV.....	77

## LISTA DE QUADROS

Quadro 1 - Palavras-chave utilizadas na pesquisa .....	18
Quadro 2 - Artigos selecionados para leitura de resumo.....	19
Quadro 3 - Reação do consumidor e suas consequências.....	23
Quadro 4 - Epícclo da análise de dados.....	27

## LISTA DE TABELAS

Tabela 1 - Número de artigos encontrados por palavras-chave. ....	19
Tabela 2 - Artigos publicados com a combinação das palavras-chave .....	19
Tabela 3 – Tunagem de hiperparâmetros. ....	68
Tabela 4 – Parâmetros Holt-Winters. ....	69
Tabela 5 – Parâmetros SARIMA. ....	69
Tabela 6 – RMSE dos métodos por item. ....	71
Tabela 7 – MAPE dos métodos por item.....	72

## LISTA DE ABREVIATURAS E SIGLAS

SKU – *Stock Keeping Unit*

ID3 - *Induction Decision Tree*

CART - *Classification and Regression Trees*

RNAs - *Redes neurais artificiais*

CNN - *rede neural convolucional*

RNN - *Recurrent Neural Networks*

LSTM - *Long Short-Term Memory*

MAE - *Mean Absolute Error*

MSE - *Mean Squared Error*

RMSE - *Root Mean Squared Error*

MAPE - *Mean absolute percentage error*

SHAP - *SHapley Additive exPlanations*

SC – *Santa Catarina*

AIC - *Aikake information criterion*

SARIMA - *Seasonal Autoregressive Integrated Moving Average*





## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO .....</b>	<b>15</b>
1.1	JUSTIFICATIVA .....	15
1.2	OBJETIVOS.....	15
<b>1.2.1</b>	<b>Objetivo Geral .....</b>	<b>15</b>
<b>1.2.2</b>	<b>Objetivos Específicos .....</b>	<b>16</b>
1.3	LIMITAÇÕES.....	16
1.4	ESTRUTURA DO TRABALHO .....	16
<b>2</b>	<b>REVISÃO DA LITERATURA E REFERENCIAL TEÓRICO .....</b>	<b>18</b>
2.1	REVISÃO BIBLIOGRÁFICA .....	18
2.2	PREVISÃO DE VENDAS.....	21
<b>2.2.1</b>	<b>Efeito da ruptura .....</b>	<b>23</b>
2.3	ANÁLISE DE DADOS .....	25
2.4	MODELOS DE INFERÊNCIA/PREDIÇÃO .....	28
<b>2.4.1</b>	<b>Regressão linear .....</b>	<b>28</b>
<b>2.4.2</b>	<b>Regressão múltipla.....</b>	<b>29</b>
<b>2.4.3</b>	<b>Suavização exponencial .....</b>	<b>29</b>
<b>2.4.4</b>	<b>Método Holt-Winters.....</b>	<b>30</b>
<b>2.4.5</b>	<b>Box-Jenkins (ARIMA).....</b>	<b>30</b>
<b>2.4.6</b>	<b>Árvore de decisão.....</b>	<b>31</b>
<b>2.4.7</b>	<b>Floresta aleatória .....</b>	<b>35</b>
<b>2.4.8</b>	<b>Gradient Boosting.....</b>	<b>37</b>
<b>2.4.9</b>	<b>XGBoost .....</b>	<b>38</b>
2.4.9.1	<i>Complexidade do modelo.....</i>	<i>42</i>
2.4.9.2	<i>Score da estrutura.....</i>	<i>42</i>
2.4.9.3	<i>Vantagens do XGBoost .....</i>	<i>45</i>

2.5	<i>CROSS VALIDATION</i> (VALIDAÇÃO CRUZADA).....	46
2.6	MÉTRICAS DE AVALIAÇÃO.....	49
2.7	SHAP (SHAPLEY ADDITIVE EXPLANATIONS).....	51
<b>3</b>	<b>PROCEDIMENTOS METODOLÓGICOS .....</b>	<b>52</b>
3.1	ENQUADRAMENTO DA PESQUISA.....	52
3.2	ROTEIRO METODOLÓGICO.....	53
<b>4</b>	<b>CONTEXTUALIZAÇÃO DO PROBLEMA.....</b>	<b>55</b>
4.1	A EMPRESA.....	55
4.2	ITENS ESCOLHIDOS PARA A ANÁLISE.....	56
<b>5</b>	<b>DESENVOLVIMENTO .....</b>	<b>60</b>
5.1	ANÁLISE EXPLORATÓRIA .....	60
<b>5.1.1</b>	<b>Análise do conjunto inicial de dados.....</b>	<b>60</b>
<b>5.1.2</b>	<b>Decomposição das séries temporais.....</b>	<b>63</b>
5.2	GERAÇÃO DE <i>FEATURES</i> .....	66
<b>5.2.1</b>	<b>Ruptura Ponderada .....</b>	<b>66</b>
<b>5.2.2</b>	<b>Variação do preço médio.....</b>	<b>66</b>
<b>5.2.3</b>	<b>Vendas em tempos passados.....</b>	<b>67</b>
<b>5.2.4</b>	<b>Média por família .....</b>	<b>67</b>
<b>5.2.5</b>	<b>Variação em relação ao ano passado.....</b>	<b>67</b>
5.3	TUNAGEM/OTIMIZAÇÃO DE HIPERPARÂMETROS.....	67
5.4	MODELOS ESTATÍSTICOS.....	68
<b>5.4.1</b>	<b>Holt-Winters .....</b>	<b>68</b>
<b>5.4.2</b>	<b>SARIMA.....</b>	<b>69</b>
<b>6</b>	<b>APRESENTAÇÃO E ANÁLISE DOS RESULTADOS.....</b>	<b>70</b>
6.1	VERIFICAÇÃO DE <i>OVERFITTING</i> .....	70
6.2	ANÁLISE GERAL DOS MODELOS.....	71
6.3	IMPORTÂNCIA DAS <i>FEATURES</i> .....	72
6.4	ANÁLISE VISUAL DAS PREVISÕES OBTIDAS PARA OS ITENS.....	75

<b>7</b>	<b>CONCLUSÕES E RECOMENDAÇÕES .....</b>	<b>78</b>
<b>8</b>	<b>REFERÊNCIAS.....</b>	<b>79</b>
	<b>APÊNDICE A – Base de dados referente ao item I .....</b>	<b>82</b>
	<b>APÊNDICE B – Base de dados referente ao item II.....</b>	<b>83</b>
	<b>APÊNDICE C – Base de dados referente ao item III .....</b>	<b>84</b>
	<b>APÊNDICE D – Base de dados referente ao item IV .....</b>	<b>85</b>



## **1 INTRODUÇÃO**

As previsões auxiliam as tomadas de decisões da cadeia de suprimentos e são fundamentais em um contexto onde o cliente anseia cada vez mais por disponibilidade de produtos e agilidade na entrega, ao mesmo tempo que há por parte da empresa a necessidade de gerenciar recursos escassos.

Garantir a disponibilidade do produto ao cliente gera elevados custos logísticos de estoque. Mas, “sem um estoque adequado, a atividade de marketing poderá detectar perdas de vendas e declínio da satisfação do cliente” (BOWERSOX e CLOSS, 2009).

As séries de vendas no varejo pertencem a um tipo especial de série temporal que normalmente contém padrões de tendência e sazonais, apresentando desafios no desenvolvimento de modelos de previsão eficazes. A análise de dados surge nos últimos anos como uma ferramenta poderosa para detectar padrões em grandes volumes de dados e extrair inteligência para guiar as organizações nas tomadas de decisões.

### **1.1 JUSTIFICATIVA**

A motivação deste trabalho originou-se das dificuldades enfrentadas no dia-a-dia no momento de prever a demanda de itens do varejo no momento da decisão de compra para o abastecimento das lojas em um horizonte curto de tempo. Aliado a isso, a crescente importância que campos como aprendizado de máquinas, análise de dados e inteligência artificial têm ganhado em mercados acirrados foi fundamental para a escolha do tema.

### **1.2 OBJETIVOS**

#### **1.2.1 Objetivo Geral**

O objetivo geral deste trabalho consiste em apresentar um modelo de previsão da demanda de itens do mercado varejista do setor de construção civil.

### 1.2.2 Objetivos Específicos

Os objetivos específicos que precisam ser atendidos para que se alcance o objetivo geral são:

- a) Criar *features* a serem utilizadas no modelo preditivo a fim de aumentar sua precisão;
- b) Aplicar metodologia de validação cruzada adaptada a séries temporais de forma a garantir a capacidade de generalização do modelo;
- c) Aplicar metodologia de tunagem/otimização de hiperparâmetros com o objetivo de tunar o modelo preditivo e obter um melhor resultado de predição;
- d) Classificar na forma de *ranking* as *features* do modelo com relação à sua relevância e impacto na predição;
- e) Comparar o modelo com outros métodos estatísticos utilizando-se de métricas de avaliação.

### 1.3 LIMITAÇÕES

O presente trabalho propõe-se a prever a demanda de itens com base no seu histórico de vendas, preço e outras características específicas do item. Informações como presença de empresas concorrentes ou exposição dos itens no ponto de venda não são levadas em consideração. A previsão foi realizada para uma amostra de quatro itens, cada um com perfil de demanda diferenciado, dentre um universo de 45 mil itens comercializados. Embora a grande maioria de itens não pôde ser analisada, pode-se inferir que a metodologia utilizada poderia ser aplicada a estes itens caso houvesse interesse e disponibilidade de tempo para tanto.

### 1.4 ESTRUTURA DO TRABALHO

O trabalho está dividido em sete capítulos. No primeiro é apresentada a introdução ao tema. A justificativa para o desenvolvimento do trabalho, os objetivos gerais e específicos além das limitações do trabalho são descritos na sequência.

No segundo capítulo é feita a revisão bibliográfica do tema e depois os principais conceitos que servem de base para o entendimento do trabalho são apresentados como referencial teórico.

O terceiro capítulo mostra o enquadramento da pesquisa e o roteiro metodológico adotado no trabalho.

A contextualização do problema com a apresentação da empresa e o detalhamento do perfil da demanda dos itens analisados estão incluídos no capítulo quatro.

O quinto capítulo consiste na análise exploratória de dados, geração de *features* e a modelagem dos métodos de previsão utilizados.

No sexto capítulo são apresentados os resultados obtidos com o modelo de previsão bem como uma análise comparativa com resultados obtidos com outros modelos estatísticos.

No sétimo e último capítulo tem-se as conclusões derivadas dos resultados obtidos e são sugeridas recomendações para trabalhos futuros.

## 2 REVISÃO DA LITERATURA E REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conceitos teóricos, que servem de base para o entendimento dos temas abordados durante o trabalho. Inicialmente é apresentada a revisão da literatura de estudos relacionados ao presente trabalho. Em seguida, são expostos os conceitos de modelos de inferência e de predição, incluindo modelos supervisionados de aprendizado de máquina - estes modernos e robustos.

### 2.1 REVISÃO BIBLIOGRÁFICA

A pesquisa bibliográfica deste trabalho tem o intuito de identificar qual o estado da arte em termos de previsão de vendas no varejo utilizando modelos de aprendizado supervisionado. A revisão bibliográfica seguiu procedimento metodológico baseado na proposta do Laboratório de Metodologias Multicritério em Apoio à Decisão da UFSC (LABMCDA). A construção do portfólio de artigos é composta por oito etapas: definição das bases de dados, definição de palavras chave, busca e filtragem na base de dados, seleção de artigos por alinhamento do título à pesquisa, seleção por reconhecimento científico, repescagem de referências excluídas, leitura de resumos e fichamento e seleção dos artigos para compor o portfólio de leitura.

As bases de dados escolhidas para o primeiro passo foram: ScienceDirect, IEEE Xplore e Google Scholar, por tratarem-se de bases conhecidas no meio da Engenharia. As palavras-chave foram divididas em dois grupos: florestas aleatórias e previsão de vendas. Para cada grupo usou-se palavras-chave em inglês na pesquisa nos bancos de dados. O Quadro 1 mostra os dois grupos e suas respectivas palavras-chave.

Quadro 1 – Palavras-chave utilizadas na pesquisa

<b>Grupo</b>	<b>Palavras-chave</b>
Floresta aleatória	- <i>Random forest</i> - <i>Gradient boosting</i> - <i>Xgboost</i>
Previsão de vendas	- <i>Sales Forecast</i> - <i>Retail</i>

Fonte: Elaborado pelo autor (2019).



Após a definição das palavras-chave, fez-se o levantamento quantitativo destas nas três bases anteriormente selecionadas. Os resultados da busca são mostrados na Tabela 1.

Tabela 1 - Número de artigos encontrados por palavras-chave.

Palavras-chave	IEEE Xplore	Science Direct	Google Scholar
<i>Random Forest</i>	5.485	151.878	2.940.000
<i>Gradient Boosting</i>	918	42.048	117.000
<i>Xgboost</i>	169	315	6.320
<i>Sales Forecast</i>	846	54.439	722.000
<i>Retail</i>	5.212	125.317	2.960.000

Fonte: Elaborado pelo autor (2019).

Analisando o resultado, observa-se que o Google Scholar apresenta mais artigos publicados, seguido pelo Science Direct, enquanto que o IEEE Xplore mostra acervo de trabalhos menor quando comparado aos dois primeiros.

O passo seguinte se dá pelo cruzamento das palavras-chave, visando refinar a busca e encontrar estudos semelhantes ao presente trabalho. A Tabela 2 contém os resultados da pesquisa dos cruzamentos entre as palavras-chave.

Tabela 2 - Artigos publicados com a combinação das palavras-chave

Palavras-chave	IEE Xplore	Science Direct	Google Scholar
<i>Random forest, sales forecast</i>	8	1.181	28.900
<i>Gradient boosting, sales forecast</i>	4	517	7.210
<i>Xgboost, sales forecast</i>	1	12	584
<i>Retail, Sales forecast</i>	34	12.539	264.000
<i>Retail, Sales forecast, Random forest</i>	1	328	24.700

Fonte: Elaborado pelo autor (2019).

Foram escolhidos alguns artigos com base na proximidade do título com o tema abordado no trabalho, para então ler os seus resumos. O Quadro 2 lista esses artigos selecionados, bem como seu autor e ano de publicação.

Quadro 2 - Artigos selecionados para leitura de resumo

<b>Título</b>	<b>Autor (Ano)</b>
<i>Sales Forecasting for Retail Chains</i>	JAIN et al. (2014)
<i>Considerations of a retail forecasting practitioner</i>	SEAMAN (2018)
<i>A big data driven framework for demand-driven forecasting with effects of marketing-mix variables</i>	KUMAR et al. (2019)
<i>A comparison of AdaBoost algorithms for time series forecast combination</i>	BARROW, CRONE (2016)
<i>Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information</i>	MA (2016)
<i>Boosting techniques for nonlinear time series models</i>	ROBINZONOV et al. (2011)

Fonte: Elaborado pelo autor (2019).

No seu artigo, Jain et al. (2014) descrevem o modelo de predição de vendas em uma importante rede de farmácia europeia. O modelo proposto utilizou um conjunto de dados econômicos e dados temporais, como vendas, promoções, competidores, feriados, localização da farmácia, acessibilidade e a época do ano. O método utilizado foi o *Extreme Gradient Boosting* e seus resultados foram comparados a outros métodos como regressão linear e florestas aleatórias, tendo resultado melhor que estes dois últimos.

Seaman (2018) discute as importantes decisões que se precisa tomar ao fazer a previsão de vendas ao nível de item. Diferentes caminhos podem ser tomados que levam a objetivos, foco e métricas de erro diferentes como, por exemplo, previsão orientada por preço ou gestão de estoque. Também menciona a questão dos *tradeoffs* inerentes aos métodos de previsão que vão impactar no resultado, tempo de execução e acurácia do modelo.

O estudo de Kumar et al. (2019) investigou a contribuição de campanhas de marketing, preço, demanda histórica e outros fatores para criar um modelo de previsão de demanda que atenda às necessidades futuras dos clientes. Segundo seus autores, a maioria dos modelos existentes são limitados pois não oferecem informações de sazonalidade e o impacto da previsão na magnitude do efeito chicote. O método utilizado é o de rede neural de propagação reversa e seus resultados superaram outros métodos de previsão.

Para Barrow e Crone (2016), o uso de algoritmos como o *AdaBoost* - variação do *Gradient Boosting* - não teve sua acurácia comprovada. Para mudar esse panorama, os autores conduziram um rigoroso estudo a partir de séries temporais provenientes de 111 indústrias e testaram o algoritmo com diferentes combinações de parâmetros. Foi observado que apenas

alguns parâmetros do *Boosting* aumentavam a acurácia, enquanto os parâmetros derivados de uma combinação de fatores de previsão geravam ganhos maiores.

Ma et al. (2016) desenvolveram uma metodologia em quatro etapas para criar um modelo de previsão que permite prever o volume de vendas de lojas de varejo em nível de SKU. Para isso são analisados os efeitos de promoções inter-categorias e intra-categorias. O método constrói variáveis explicativas e seleciona as que melhor descrevem o comportamento das vendas. As previsões desse modelo tiveram melhor resultado quando comparadas a modelos que não levam em conta as variáveis explicativas.

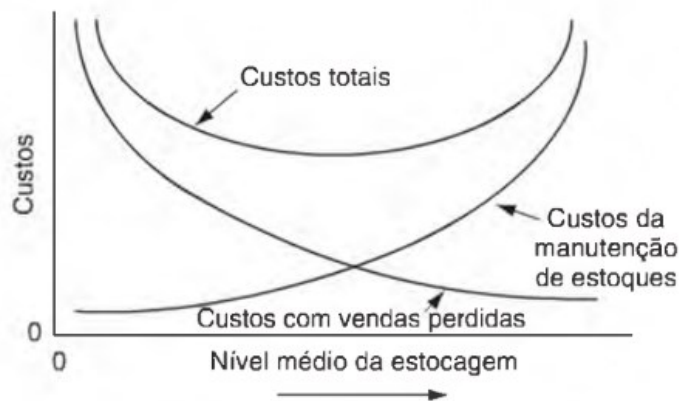
Robinsonov et al. (2012) propõem um modelo linear e aditivo para identificar relações não-lineares entre os dados de séries temporais. Um algoritmo de *gradient boosting* é aplicado para simultaneamente ajustar os modelos, selecionar as variáveis e escolher o melhor modelo. Além disso, adicionaram variáveis exógenas que ajudaram a melhorar os resultados da previsão, que foram superiores aos resultados de modelos que não se utilizavam dessas variáveis.

## 2.2 PREVISÃO DE VENDAS

As previsões auxiliam na tomada de decisão relativa a itens da cadeia de suprimentos e são fundamentais em um contexto onde o cliente anseia cada vez mais por disponibilidade de produtos e agilidade na entrega, ao mesmo tempo que há por parte da empresa a necessidade de gerenciar recursos escassos.

Garantir a disponibilidade do produto ao cliente gera elevados custos logísticos de estoque. Mas, “sem um estoque adequado, a atividade de marketing poderá detectar perdas de vendas e declínio da satisfação do cliente” (BOWERSOX e CLOSS, 2009). A Figura 1 mostra o *trade-off* entre custos da manutenção de estoques e custo com vendas perdidas que influencia decisões estratégicas de gestão de estoques no varejo. Com o aumento dos níveis médios de estoque, há maior disponibilidade de produtos aos clientes e, por consequência, diminui-se o custo com vendas perdidas. Entretanto, esta política acarreta em custos maiores de manutenção do estoque.

Figura 1 - Trade-off custos x nível de estoque



Fonte: Ballou (2006).

Tendo em vista que o consumidor não está disposto a pagar a mais por ter o produto disponível no momento e local certo, tem-se a necessidade de estabelecer o equilíbrio entre nível de serviço e custo. Para que a organização tenha uma gestão eficaz e eficiente do negócio, é preciso que se tenha suporte informacional que permita tomar decisões baseadas nessas informações de forma a responder rapidamente a alterações de demanda por parte dos consumidores.

Segundo Hanke (2009), os métodos de previsão de venda formais envolvem estender as experiências do passado para o futuro. Supõe-se que as condições que geraram relações e dados passados são indistinguíveis das condições do futuro. A previsão não pode ser mais precisa do que os dados em que se baseia. Existem quatro critérios que podem ser aplicados para determinar se os dados serão úteis (HANKE, 2009):

- I. Os dados devem ser confiáveis e precisos. Deve-se tomar o cuidado adequado para que os dados sejam coletados de uma fonte confiável com a devida atenção à precisão;
- II. Os dados devem ser relevantes. Os dados devem ser representativos das circunstâncias para as quais eles estão sendo usados;
- III. Os dados devem ser consistentes;
- IV. Os dados devem ser de um horizonte de tempo adequado. Não se pode ter poucos dados (não há história suficiente na qual basear os resultados futuros) ou muitos dados (dados de períodos históricos irrelevantes no passado).

As séries de vendas no varejo pertencem a um tipo especial de série temporal que normalmente contém padrões de tendência e sazonalidade apresentando desafios no desenvolvimento de modelos de previsão eficazes.

### 2.2.1 Efeito da ruptura

A ruptura de estoque é entendida como a falta do produto na gôndola, no momento em que o consumidor deseja comprar. A ruptura ocorre quando a demanda real é maior do que a demanda que foi prevista. O custo associado à falta de produtos nas prateleiras é difícil de mensurar, porque difere em função da resposta do consumidor (CRUZ, 2015). Gruen, Corsten e Bharadwaj (2002) buscaram medir a resposta do consumidor à ruptura a partir das seguintes possibilidades: substituir o item dentro da mesma marca, comprar o item em outra marca, comprar o item em outra loja, adiar a compra ou simplesmente não comprar item nenhum. As consequências de cada resposta do consumidor estão listadas no Quadro 3.

Quadro 3 – Reação do consumidor e suas consequências.

Atitude	Varejista	Fabricante
Substituir o item pela mesma marca	Risco do consumo de produtos de menor valor	Sem prejuízos financeiros, risco do consumo de produtos de menor valor
Compra o item por outra marca	Risco do consumo de produtos de menor valor	Perda de receita, desgaste da marca e risco de fidelização ao concorrente
Compra o item em outra loja	Perda de receita, desgaste da imagem da loja	Desgaste da imagem da marca
Desistir	Perda da receita e desgaste da imagem	Desgaste da marca, perda da receita e de fidelidade
Adiar a Compra	Sem prejuízos financeiros, apenas de imagem	Sem prejuízos financeiros, apenas de imagem da marca

Fonte: Gruen et al. (2002).

Para o varejista, foco do presente estudo, a substituição do item pela mesma marca ou outra marca acarreta no risco de o consumidor optar por um produto de menor valor, diminuindo assim sua receita e passando a sensação de que não pode-se contar com a disponibilidade dos produtos de sua preferência. O ato de adiar a compra não traz prejuízos financeiros, somente os prejuízos de imagem. Os casos mais graves são quando o consumidor desiste da compra ou compra o item em outra loja. Nessas situações a loja tem tanto o prejuízo financeiro quanto o desgaste da imagem pelo consumidor.

Qualquer que seja a reação do consumidor, a ruptura do item é negativa para o fabricante. Os piores casos são quando o cliente desiste da compra ou opta por comprar um item de um fabricante concorrente, sendo este mais grave pois corre o risco de o cliente fidelizar-se ao concorrente.

O resultado da pesquisa (Figura 2) mostra que em 43% dos casos há grandes prejuízos financeiros para o varejista (desistir da compra e comprar item em outra loja) e em 31% dos casos há prejuízo para o fabricante (substituir o item por outra marca e desistir da compra).

Figura 2 – Resposta do consumidor à ruptura.



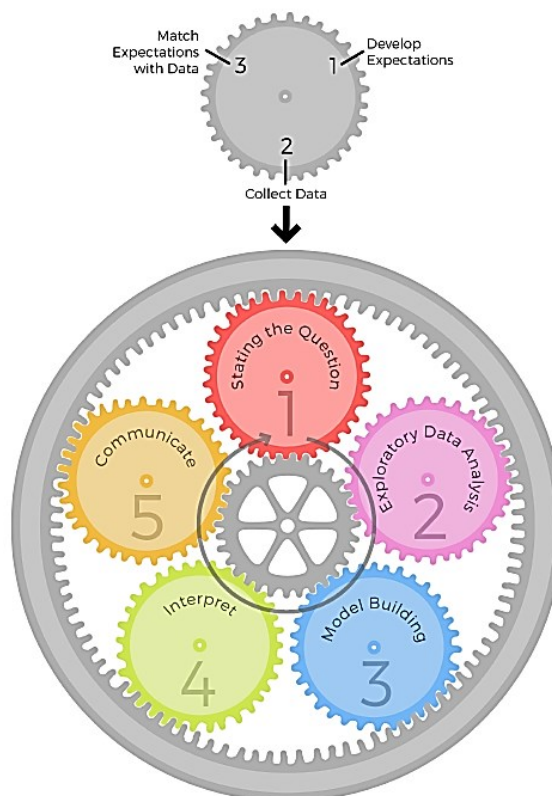
Fonte: Gruen et al. (2002)

Por esses motivos, a ruptura é um fenômeno que deve ser evitado ao máximo pelas lojas varejistas e o equilíbrio entre manter altos níveis de estoque e o nível de serviço desejado é um *trade-off* fundamental na estratégia de empresas situadas em um mercado tão competitivo como o do varejo de materiais de construção.

## 2.3 ANÁLISE DE DADOS

Cada problema de análise de dados apresenta suas particularidades e, portanto, deve ser resolvido de maneira única. Diferente de outros processos, a análise de dados não é um processo linear onde segue-se um passo-a-passo que no final resultará em uma resposta pronta e coerente. Trata-se de um processo altamente iterativo onde informação é aprendida a cada etapa, que então informa quando e como devemos refinar o passo anterior, ou como proceder nas próximas ações. Peng e Matsui (2016) descreveram esse processo como o epiciclo da análise de dados. O epiciclo é um pequeno ciclo composto por três etapas, cujo centro se move sobre a circunferência de um ciclo maior, composto por cinco macroetapas, conforme ilustrado na Figura 3.

Figura 3 - Epiciclo da análise de dados.



Fonte: Peng e Matsui (2016).

Segundo Peng e Matsui (2016), as cinco macroetapas são:

- I. Estabelecimento e refinamento da questão;
- II. Exploração de dados;

- III. Construção de modelos formais estatísticos;
- IV. Interpretação de resultados;
- V. Comunicação de resultados.

Em cada uma dessas macroetapas é fundamental que se verifiquem as três etapas interativas do epícciclo, as quais são:

- I. Estabelecer expectativas: Consiste em pensar no que se espera obter da análise, antes mesmo de fazer qualquer inspeção ou teste com os dados;
- II. Coletar informações e comparar com as expectativas: Essa etapa engloba a coleta de informação sobre a questão ou os próprios dados. Para a questão podem ser feitas pesquisas na literatura ou consulta à especialistas para assegurar que a questão é boa. Quanto aos dados, devem ser coletados conforme as expectativas estabelecidas;
- III. Revisar expectativas ou revisar os dados de forma que eles estejam alinhados: Caso as expectativas estejam compatíveis com os dados, ótimo. Do contrário, tem-se duas situações: ou é necessário revisar as expectativas ou corrigir os dados.

O ciclo maior do epícciclo é formado quando se é aplicado repetitivamente essas etapas durante as cinco macroetapas que compõem uma análise de dados completa, conforme descrito no Quadro 4.

Quadro 4 - Epícciclo da análise de dados.

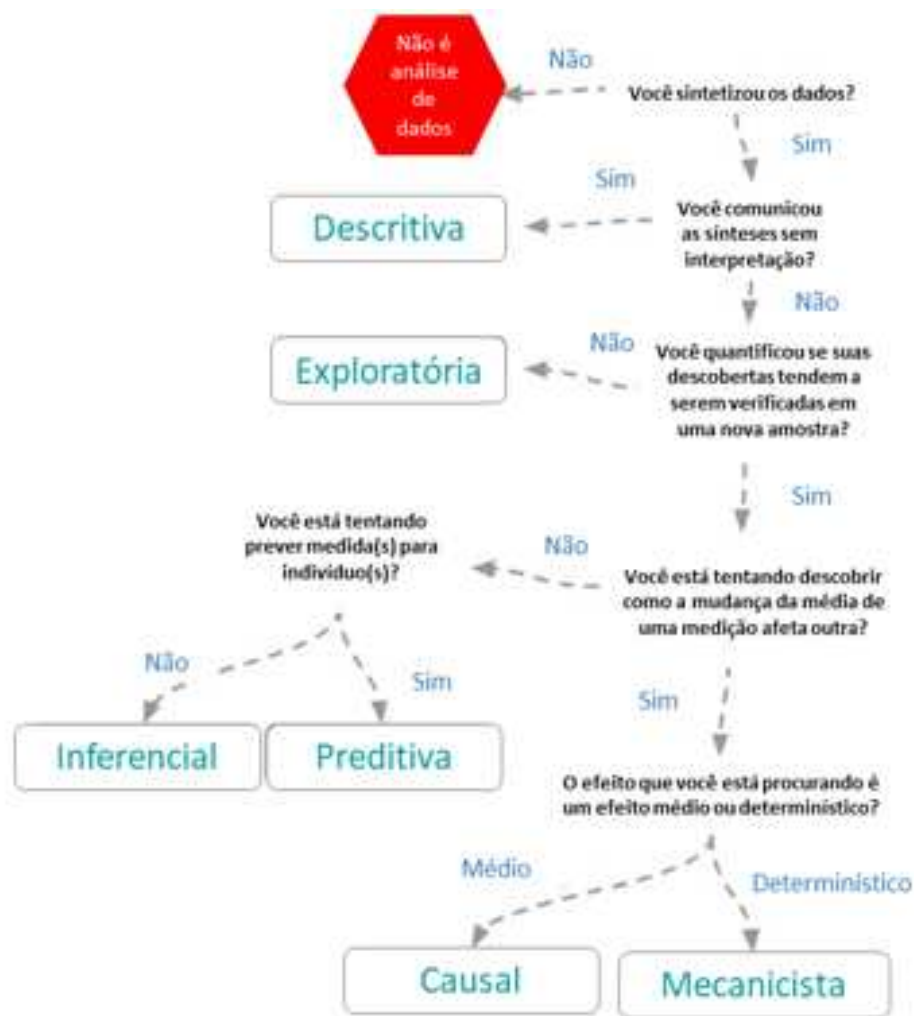
Macroetapas	Expectativa	Coleta de Informações	Revisão de Expectativas
Estabelecendo a Questão	A questão é de interesse da audiência	Busca na Literatura/Especialistas	Detalhar questão contundente
Exploração dos Dados	Dados são apropriados para a questão	Fazer gráficos exploratórios de dados	Refinar a questão ou coletar dados adicionais
Modelagem Formal	Modelo primário responde a questão	Formular modelos secundários e/ou prover análises de sensibilidade	Revisar modelo formal para incluir novos preditores
Interpretação dos Resultados	Interpretação da análise provê uma resposta específica e contundente para a questão	Interpretar a totalidade das análises com foco na magnitude dos efeitos e incerteza	Revisar a exploração de dados e/ou modelos para prover respostas específicas e interpretáveis
Comunicação	Processo e resultados da análise são entendidos, completos e significativos para a audiência	Buscar <i>feedbacks</i>	Revisar análises ou abordagem para apresentação

Fonte: Peng e Matsui (2016). Traduzido.



Em análise de dados, deve-se definir o caráter da questão antes mesmo de iniciar a análise propriamente dita. Isso dá-se pelo fato de o tipo da questão escolhida influenciar na estratégia a ser adotada na modelagem do modelo. Leek (2015) afirma que algumas questões são mais fáceis de responder com dados e outras são mais difíceis, criando então uma categorização dos tipos de questões em análises de dados pela facilidade de responder as questões com dados. A Figura 4 mostra o fluxograma proposto por Leek.

Figura 4 - Fluxograma dos tipos de questões em análise de dados.



Fonte: Leek (2015). Adaptado.

A questão descritiva é aquela que procura sumarizar uma característica da base de dados, sem qualquer interpretação do resultado. A questão exploratória busca analisar a base de dados na busca por padrões, tendências ou relações entre variáveis. É também chamada de

“geradora de hipóteses”.

Especificamente para o presente estudo, faz-se importante definir se a questão respondida será inferencial ou preditiva.

A questão inferencial tem o objetivo de estimar uma associação entre um preditor de interesse e o resultado, ajustando essa associação para qualquer variável *confounding*. Normalmente a questão inferencial apresenta um número reduzido de preditores, mas com muitas variáveis *confounding* a se considerar. São feitas então análises de sensibilidade para ver se as associações de interesse são robustas para diferentes conjuntos de *confounders*. Segundo Wayne (2016), uma variável *confounding* é aquela que tem relação tanto com a variável preditora quanto com o resultado, podendo causar distorções na associação entre estes.

As questões preditivas têm como objetivo identificar o modelo que melhor prediz o resultado, sem importar-se com os preditores, *confounders* e de que forma eles influenciam no resultado. Deseja-se apenas desenvolver um modelo que prediz os resultados com o menor erro possível.

A questão causal é aquela que procura saber se a alteração de um fator irá provocar alteração em outro fator, na média, de uma população. Questões causais são de difícil predição em ambientes experimentais pouco controlados.

E por fim, a questão mecanicista é aquela que busca resposta de “como” a alteração de um fator irá causar alteração em outro fator, na média, de uma população. Por exemplo, décadas de dados mostram uma clara relação causal entre fumar e câncer.

## 2.4 MODELOS DE INFERÊNCIA/PREDIÇÃO

### 2.4.1 Regressão linear

A regressão linear simples busca relacionar o comportamento de uma variável Y a ser prevista, chamada de variável dependente, com uma variável explicativa X, chamada variável independente. A Equação 1 representa o cálculo da regressão simples.

$$Y = a + bX + e \tag{1}$$

onde *a* e *b* são coeficientes que devemos calcular; *e* representa o erro de previsão, ou seja, a diferença entre o valor real e o valor previsto. Para estimar os valores de *a* e *b*, é aplicado o

método dos mínimos quadrados, que busca minimizar a soma dos erros quadrados, evitando assim o problema de somar valores positivos e negativos.

Na prática, esse tipo de regressão com apenas duas variáveis é pouco usado, dado que no mundo real as situações quase sempre envolvem relações entre mais de duas variáveis.

### 2.4.2 Regressão múltipla

A regressão múltipla, como o nome sugere, busca determinar o valor da variável  $Y$  pela relação linear com duas ou mais ( $k$ ) variáveis independentes.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + e \quad (2)$$

Da mesma forma que acontece com a regressão simples, devemos estimar todos coeficientes  $b_k$  de forma a minimizar a soma dos erros ao quadrado da reta de regressão.

### 2.4.3 Suavização exponencial

Esse método consiste em prever a variável desejada através das médias ponderadas de observações antigas, atribuindo um peso que decresce exponencialmente à medida que os dados ficam mais antigos. Ou seja, os dados mais recentes têm maior prioridade no cálculo. Tal priorização ocorre com o uso da constante de suavização  $\alpha$ , que deve satisfazer a condição de  $0 < \alpha < 1$ . Esse parâmetro determina a taxa na qual a influência das observações passadas decai exponencialmente. Um valor perto de 1 indica rápido aprendizado (ou seja, o dado mais recente é quem influencia a previsão), enquanto um valor de  $\alpha$  perto de 0 indica aprendizado lento (observações passadas tem maior influência nas previsões). A fórmula da suavização exponencial (Equação 3) é dada por:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \quad (3)$$

onde:

$\hat{Y}_{t+1}$  = o novo valor suavizado ou a previsão para o próximo período

$\alpha$  = constante de suavização  $Y_t$

$Y_t$  = a nova observação ou o real valor da série no período t

$\hat{Y}_t$  = o valor antigo suavizado ou a previsão do período t

A suavização exponencial é um método utilizado para se obter previsões confiáveis a curto prazo em séries temporais que não apresentam tendência ou sazonalidade.

#### 2.4.4 Método Holt-Winters

Esse modelo foi criado para realizar a previsão de séries temporais com tendência e sazonalidade, que consistem em fatores que aparecem com frequência quando se trabalha com dados reais. Para tal, dois coeficientes são adicionados ao método anterior. O coeficiente  $b_t$  é responsável por captar as tendências de crescimento ou queda e representa a inclinação da reta dessa tendência. A sazonalidade é representada pela determinação do coeficiente  $\gamma$ , que é o índice de sazonalidade. Os cálculos para o método Holt-Winters estão apresentados nas Equações 4, 5, 6 e 7:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (4)$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \quad (5)$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-s} \quad (6)$$

$$F_{t+m} = (L_t + mb_t)S_{t-s+m} \quad (7)$$

O objetivo do método é determinar os coeficientes,  $\alpha$ ,  $b_t$  e  $\gamma$  que otimizam a previsão, ou seja, que apresentem menor erro.

#### 2.4.5 Box-Jenkins (ARIMA)

O método de Box-Jenkins é um modelo auto-regressivo integrado de médias móveis (autoregressive integrated moving average ou ARIMA, na sigla em inglês), o qual consiste de uma generalização de um modelo auto-regressivo de médias móveis (ARMA), popular na

Estatística e Econometria, particularmente na análise de séries temporais. Ambos os modelos são ajustados aos dados da série temporal para entender melhor os dados ou para prever pontos futuros na série. Modelos ARIMA são aplicados em alguns casos em que os dados mostram evidências de não estacionariedade, em que um passo inicial de diferenciação (correspondente à parte "integrada" do modelo) pode ser aplicado uma ou mais vezes para eliminar a não estacionariedade.

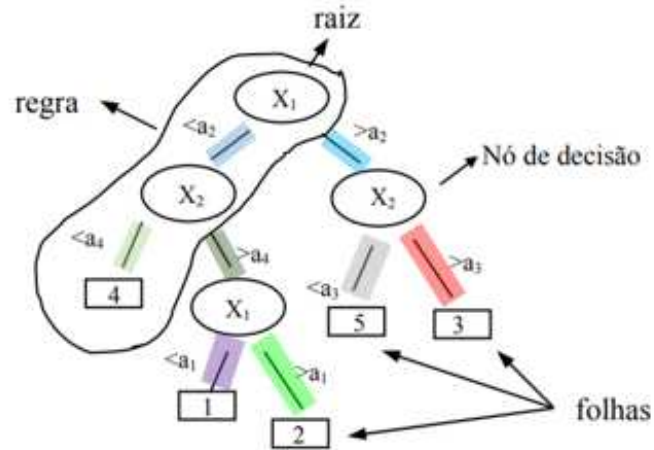
A parte auto-regressiva (AR) do modelo ARIMA indica que a variável de interesse é regressada em seus próprios valores defasados, isto é, anteriores. A parte de média móvel (MA) indica que o erro de regressão é na verdade uma combinação linear dos termos de erro, cujos valores ocorreram contemporaneamente e em vários momentos no passado. A parte integrada (I) indica que os valores de dados foram substituídos com a diferença entre seus valores e os valores anteriores e este processo diferenciador pode ter sido realizado mais de uma vez. O propósito de cada uma destas características é fazer o modelo se ajustar aos dados da melhor forma possível.

Modelos ARIMA não sazonais são geralmente denotados como  $ARIMA(p,d,q)$  em que os parâmetros  $p$ ,  $d$  e  $q$  são números inteiros não negativos,  $p$  é a ordem (número de defasagens) do modelo auto-regressivo,  $d$  é o grau de diferenciação (o número de vezes em que os dados tiveram valores passados subtraídos) e  $q$  é a ordem do modelo de média móvel. Modelos SARIMA são geralmente denotados como  $SARIMA(p,d,q)(P,D,Q)m$  em que  $m$  se refere ao número de períodos que define a sazonalidade e  $P$ ,  $D$  e  $Q$  se referem aos termos de auto-regressão, diferenciação e média móvel para a parte sazonal do modelo SARIMA.

#### 2.4.6 Árvore de decisão

A árvore de decisão é uma ferramenta de aprendizado de máquina supervisionado não-paramétrico, utilizado em problemas de classificação e regressão. A estrutura da árvore é formada por um conjunto de elementos que armazenam informações chamados nós. O nó ponto de partida da árvore é chamado raiz e possui o maior nível hierárquico, dele saem ramificações chamadas de filhos. O nó que não tem filhos é chamado folha. Esses elementos da árvore de decisão podem ser observados na Figura 5.

Figura 5 - Representação de uma árvore de decisão.

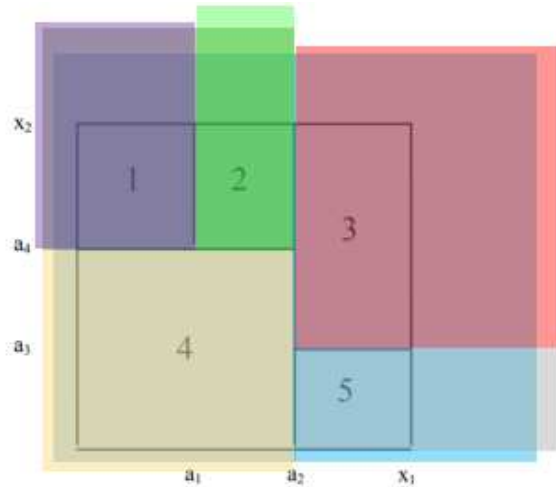


Fonte: Gama (2004).

O nó nada mais é do que um teste para algum atributo, onde seus ramos descendentes correspondem aos possíveis valores desse atributo. Cada percurso da árvore, da raiz à folha corresponde à uma regra de classificação e cada folha está associada a uma classe.

A escolha do atributo a ser testado em cada nó é baseada no ganho de informação, que é a informação aprendida ao dividir uma região do espaço em duas sub-regiões. O objetivo é conseguir o maior ganho de informação com esse atributo e o processo é aplicado para os próximos nós até que um critério de parada seja alcançado, como profundidade ou número de folhas, ou até que todas as folhas sejam puras (profundidade máxima). A Figura 6 mostra a representação no espaço da árvore de decisão.

Figura 6 - Representação no espaço de uma árvore de decisão.



Fonte: Gama (2004).

Nos casos em que a árvore é usada para classificação, os critérios para quantificar o ganho de informação mais conhecidos são baseados na Entropia e Índice Gini. A entropia é uma medida oriunda da teoria da informação que mede a falta de homogeneidade observada dos dados. Na escolha de um atributo a ser testado busca-se dividir o conjunto de dados em duas sub-regiões o mais heterogêneas possíveis. No caso onde a divisão resulta em duas regiões totalmente heterogêneas, a entropia é máxima (igual a 1) e as regiões são ditas puras.

Dado um conjunto de entrada ( $S$ ) que pode ter  $c$  classes distintas, a entropia de  $S$  será dada por:

$$Entropia(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (8)$$

onde:  $p_i$  é a proporção de dados em  $S$  que pertencem à classe  $i$ .

Dado um atributo  $A$  do conjunto de dados  $S$ , o ganho de informação representa a medida da diminuição da entropia esperada quando utiliza-se esse atributo  $A$  para dividir o conjunto de dados.

Seja:  $P(A)$  o conjunto dos valores que  $A$  pode assumir,  $x$  um elemento desse conjunto e  $S_x$

o subconjunto de  $S$  formado pelos dados em que  $A = x$

A entropia gerada ao se dividir o conjunto  $S$  quanto ao atributo  $A$  dada por:

$$E(A) = \sum_{x \in P(A)} \frac{|S_x|}{|S|} \quad (9)$$

O ganho de informação é então calculado por:

$$\text{Ganho}(S, A) = \text{Entropia}(S) - E(A) \quad (10)$$

No processo de escolha do atributo a ser testado, o atributo que representar o maior ganho de informação será o escolhido para o nó em questão, pois será que o dividirá melhor o conjunto de dados, isto é, será o atributo no qual a árvore de decisão “aprenderá” mais.

Desenvolvido por Conrado Gini (1912), o índice Gini também é utilizado para medir o grau de heterogeneidade dos dados, ou a “impureza” de um nó da árvore. Em um determinado nó o índice é calculado por:

$$\text{Índice Gini} = 1 - \sum_{i=1}^c p_i^2 \quad (11)$$

onde: " $p_i$ " é a frequência relativa de cada classe em cada nó e " $c$ " é o número de classes. Quando o Índice de Gini aproxima-se de 1, o nó é dito impuro, enquanto que quando o valor dele é zero, o nó é dito puro. Portanto, quando o Índice de Gini é usado como critério de escolha de atributo, procura-se o atributo com menor valor, pois ele vai fornecer um ganho de informação maior ao modelo.

Os algoritmos de construção de árvores de decisão mais conhecidos são o ID3 (*Induction Decision Tree*) e o CART (*Classification and Regression Trees*). O ID3 é um algoritmo desenvolvido por Ross Quinlán em 1979 que a partir do conceito de entropia, seleciona a *feature* que traz o maior ganho de informação. Quando aplicado em amostras pequenas seus resultados podem ser acometidos de *overfitting*. Também, o modelo não consegue manipular atributos numéricos, valores faltantes e apenas um atributo é testado de cada vez na tomada de decisão.

Criado em 1984 por Breiman *et al.*, o CART usa do Índice Gini para construir árvores binárias. A vantagem desse método é que ele pode manipular atributos numéricos e categóricos e lidar facilmente com valores faltantes e outliers.

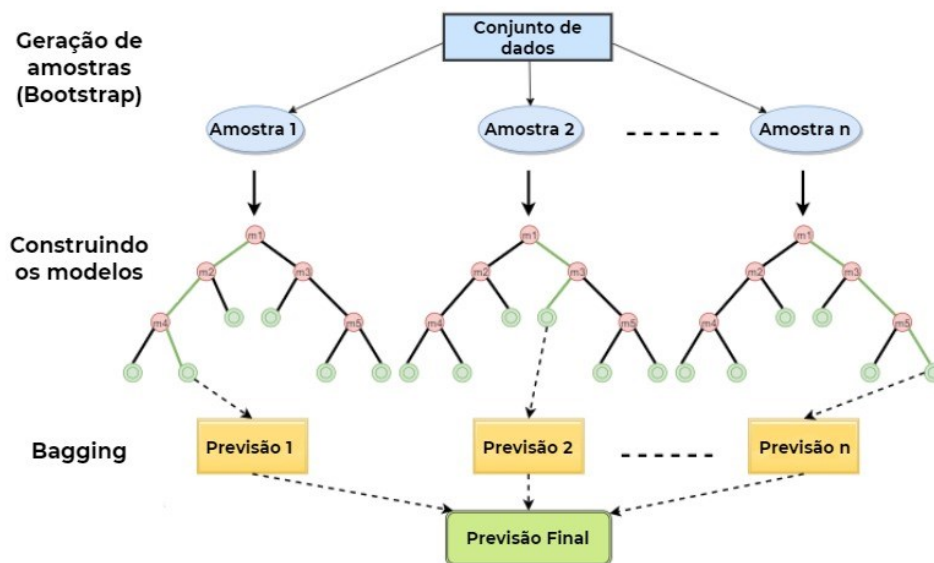


### 2.4.7 Floresta aleatória

A floresta aleatória é o resultado do crescimento de muitas árvores de decisão. É um método de aprendizagem de máquina versátil que permite tratar problemas de redução de dimensionalidade, valores faltantes e *outliers*. Nele, um grupo de modelos fracos são combinados para formar um modelo mais forte. Por isso, diz-se que é um método de aprendizagem de *ensemble*, onde modelos preditivos são agrupados de modo a melhorar a precisão e a estabilidade do modelo.

Na maioria das vezes o método de ensemble utilizado na floresta aleatória é o *Bagging*, termo introduzido em 1994 por Leo Breiman, que é uma palavra artificial formada pela combinação das palavras *bootstrap aggregating*. A técnica do *Bagging* combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados, reduzindo assim a variância das previsões. A representação do *Bagging* é apresentada na Figura 7.

Figura 7 - Representação esquemática de *Bagging* em floresta aleatória.



Fonte: Dmitrievski (2018)

Para classificar um novo objeto baseado em atributos, cada árvore dá uma classificação, como se votassem para determinada classe. A floresta escolhe a classificação que tiver mais votos (de todas as árvores da floresta) e, em caso de regressão, considera a média das saídas por árvores diferentes.

Cada árvore é plantada e cultivada da seguinte forma:

- I. Supondo que o número de casos no conjunto de treinamento é  $N$ . Então, a amostra desses  $N$  casos é escolhida aleatoriamente, mas com substituição. Esta amostra será o conjunto de treinamento para o cultivo da árvore;
- II. Se houver  $M$  variáveis de entrada, um número  $m < M$  é especificado de modo que, em cada nó,  $m$  variáveis de  $M$  sejam selecionadas aleatoriamente. A melhor divisão nestes  $m$  é usada para dividir o nó. O valor de  $m$  é mantido constante enquanto crescemos a floresta;
- III. Cada árvore é cultivada na maior extensão possível e não há poda;
- IV. Preveja novos dados agregando as previsões das árvores (ou seja, votos majoritários para classificação, média para regressão).

O uso de *bootstrap aggregating* reduz o erro quadrático médio e diminui a variância do classificador treinado. O erro não será muito diferente em diferentes amostras. Como resultado, o modelo sofrerá menos *overfitting*. A eficácia do *bagging* reside no fato de que os algoritmos básicos (árvores de decisão) são treinados em várias amostras aleatórias e seus resultados podem variar muito, enquanto que seus erros são mutuamente compensados na votação.

O algoritmo da Floresta Aleatória mostrou-se extremamente eficaz, capaz de resolver problemas práticos. Ele fornece um treinamento de alta qualidade com um número aparentemente grande de aleatoriedade, introduzido no processo de construção do modelo. O algoritmo também é eficaz em estimar dados faltantes e mantém a precisão quando grande parte dos dados estão faltando. Além disso, os erros em conjuntos de dados onde as classes são desequilibradas são equilibrados pelo modelo.

A floresta aleatória envolve a amostragem dos dados de entrada com substituição chamada como amostragem de *bootstrap*. Aqui um terço dos dados não é usado para treinamento e pode ser usado para testes. Estes são chamados de amostras de fora da cesta. O erro estimado nas amostras de fora da cesta é conhecido como erro de fora da cesta. O estudo de estimativas do erro de fora da cesta fornece evidências para mostrar que a estimativa de fora da cesta é tão precisa quanto usar um conjunto de teste do mesmo tamanho que o conjunto de treinamento. Portanto, usar a estimativa de erro de fora da cesta remove a necessidade de ter um conjunto de teste extra.

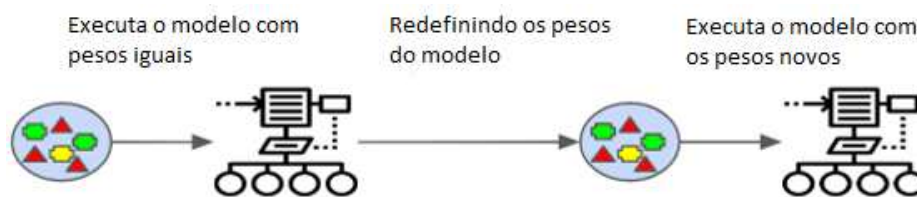
Outro benefício da floresta aleatória é o poder de lidar com dados em grandes volumes e com muitas dimensões. Ele pode lidar com milhares de variáveis de entrada e identificar as variáveis mais significativas, sendo por isso considerado um dos métodos de redução de dimensões. Além disso, o modelo produz o grau de importância das variáveis, o que ajuda a compreender a influência delas no modelo.

#### 2.4.8 Gradient Boosting

O termo *Boosting* refere-se a um grupo de algoritmos que utilizam médias ponderadas para tornar resultados de aprendizagem fraca em aprendizagem mais forte. Diferentemente do *Bagging*, onde cada modelo é executado independentemente e, em seguida, agrega-se as saídas no final sem preferência por nenhum modelo, *Boosting* (impulsionando) é tudo sobre “trabalho em equipe”. Cada modelo executado determina os recursos nos quais o próximo modelo se concentrará.

*Boosting* também requer *bootstrapping*. No entanto, ao contrário do *Bagging*, aumenta os pesos de cada amostra de dados. Isso significa que algumas amostras serão executadas com mais frequência do que outras. Quando *Boosting* executa cada modelo, ele rastreia quais amostras de dados são mais bem-sucedidas e quais não são. O esquema do *boosting* é mostrado na Figura 8.

Figura 8 - Esquema do *Boosting*.



Fonte: Ben Rogojan (2017).

Os conjuntos de dados com as saídas de maior erro recebem pesos maiores. Esses são considerados dados que têm mais complexidade e exigem mais iterações para treinar adequadamente o modelo, ou seja, as taxas de erro do modelo são monitoradas, pois os modelos melhores recebem pesos melhores.

### 2.4.9 XGBoost

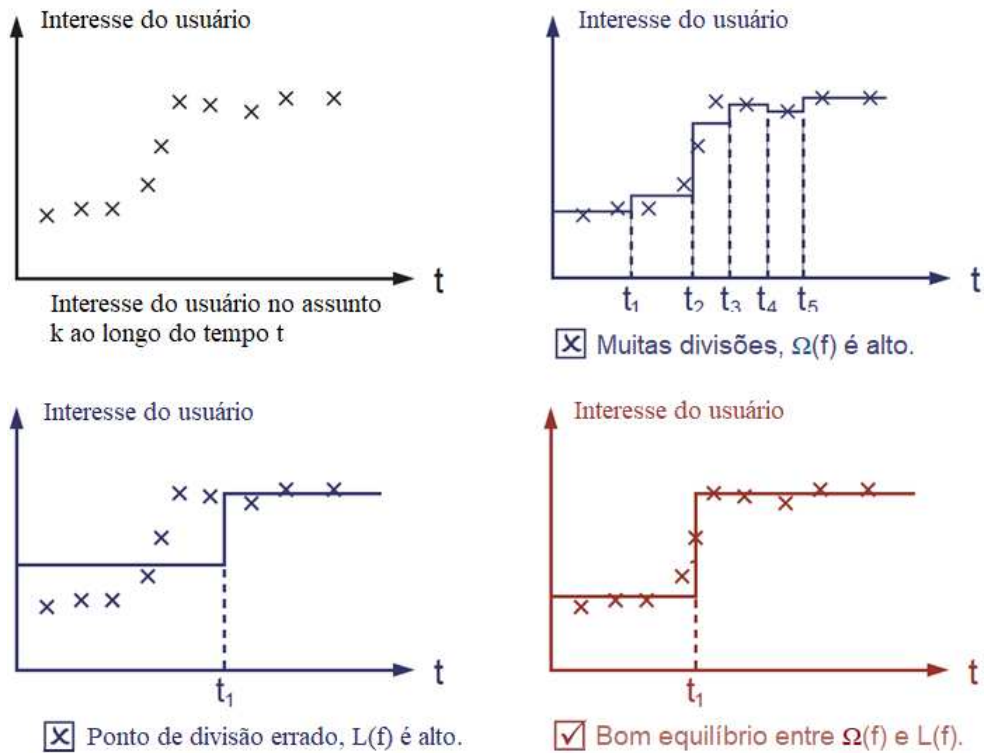
Um modelo de aprendizagem supervisionada geralmente refere-se à estrutura matemática pela qual a previsão  $y_i$  é feita a partir da entrada  $x_i$ . Os parâmetros são a parte indeterminada que precisa-se aprender com os dados. A tarefa de treinar o modelo equivale a encontrar os melhores parâmetros  $\theta$  que melhor se ajustam aos dados de treinamento  $x_i$  e a variável a ser predita  $y_i$ . Para treinar o modelo, precisa-se definir a função objetiva para medir a adequação do modelo aos dados de treinamento.

Uma característica recorrente nas funções objetivas é que elas consistem em duas partes: o erro de treinamento e o termo de regularização dada pela equação XX:

$$obj(\theta) = L(\theta) + \Omega(\theta) \quad (12)$$

onde:  $L$  é a função do erro de treinamento e  $\Omega$  é o termo de regularização. O erro de treinamento mede como o modelo é preditivo em relação aos dados de treinamento. O termo de regularização controla a complexidade do modelo, o que ajuda a evitar o *overfitting*. A Figura 9 exemplifica os dois conceitos apresentados, onde é dado uma série temporal que representa o interesse de usuários em determinado assunto.

Figura 9 – Interesse de usuários ao longo do tempo

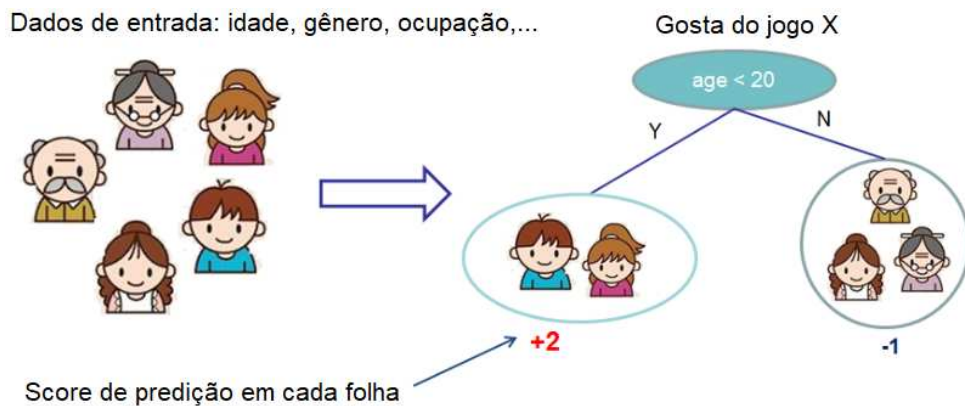


Fonte: Chen (2016). Traduzido.

A opção em vermelho é a que apresenta o melhor equilíbrio entre o erro de treinamento e o termo de regularização. O princípio geral é que se deseja um modelo simples e preditivo. O *trade-off* entre os dois também é referido como *trade-off* de viés-variância em aprendizagem de máquina.

O modelo utilizado pelo *XGBoost* é o *Decision tree ensembles*, que consiste num conjunto de árvores de classificação e regressão (CART). A Figura 10 mostra um exemplo simples de um CART que classifica se alguém vai gostar de um hipotético jogo de computador X.

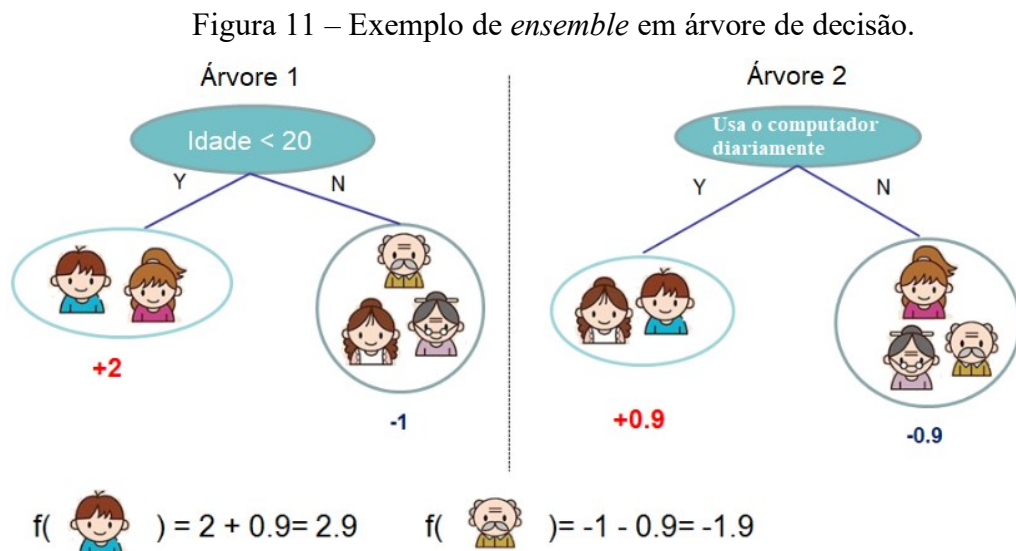
Figura 10 – Exemplo árvore de decisão.



Fonte: Chen (2016). Traduzido.

Os membros de uma família são classificados em folhas diferentes e são atribuídos a pontuação da folha correspondente. O CART é um pouco diferente das árvores de decisão, nas quais a folha contém apenas valores de decisão. No CART, uma pontuação real é associada a cada uma das folhas, o que dá interpretações mais ricas que vão além da classificação. Isso também permite uma abordagem unificada e baseada em princípios para a otimização.

Normalmente, uma única árvore não é suficientemente forte para ser usada na prática. O que é realmente usado é o modelo de *ensemble*, que soma a previsão de várias árvores juntas.



Fonte: Chen (2016). Traduzido.

Na Figura 11 tem-se um exemplo de um conjunto de árvores de duas árvores que se complementam. As pontuações de previsão de cada árvore individual são somadas para obter a pontuação final. Matematicamente, pode-se escrever o modelo na forma:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (13)$$

onde  $K$  é o número de árvores,  $f$  é uma função no espaço funcional  $F$ , e  $F$  é o conjunto de todas as possíveis CARTs. A função objetivo a ser otimizada é dada por:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \quad (14)$$

A estrutura de aprendizado de árvores é mais difícil que o problema tradicional de otimização onde pode-se simplesmente usar o gradiente. É inviável aprender todas as árvores de uma só vez. Em vez disso, usa-se uma estratégia aditiva: corrigir que foi aprendido e adicionar uma nova árvore de cada vez. A variável  $\hat{y}_i^{(t)}$  é o valor predito no passo  $t$ :

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_1) = \hat{y}_i^{(0)} + f_1(x_1) \\ \hat{y}_i^{(2)} &= f_1(x_1) + f_2(x_2) = \hat{y}_i^{(1)} + f_2(x_2) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \quad (15)$$

Considerando o MSE (*Mean Squared Error*) como a função de erro de treinamento, a função objetiva torna-se:

$$\begin{aligned} obj^t &= \sum_{i=1}^n (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^t \Omega(f_i) \\ &= \sum_{i=1}^n [2(\hat{y}_i^{(t-1)} - y_i)f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + \text{constante} \end{aligned} \quad (16)$$

A forma do MSE é amigável, com um termo de primeira ordem (geralmente chamado de residual) e um termo quadrático. Para outras perdas de interesse (por exemplo, perda logística), não é tão fácil obter uma forma tão agradável. Assim, no caso geral, utiliza-se a expansão de Taylor da função de erro de treinamento até à segunda ordem:

$$obj^t = \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + \text{constante} \quad (17)$$

Onde  $g_i$  e  $h_i$  são definidos como:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (18)$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (19)$$

Após remover todas as constantes, a função objetiva do passo  $t$  fica:

$$\sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (20)$$

Este torna-se o objetivo de otimização para a nova árvore. Uma vantagem importante desta definição é que o valor da função objetivo depende apenas de  $g_i$  e  $h_i$ . É assim que o XGBoost suporta funções de perda personalizadas. Pode-se otimizar todas as funções de perda, incluindo regressão logística e ranking em pares, usando exatamente o mesmo resolvidor que usa  $g_i$  e  $h_i$  como input.

#### 2.4.9.1 Complexidade do modelo

O termo de regularização serve para controlar a complexidade do modelo e para isso é preciso definir a complexidade da árvore  $\Omega(f)$ . A definição de uma árvore  $f(x)$  é:

$$f_t(x) = w_{q(x)}, w \in R^t, q: R^d \rightarrow \{1, 2, \dots, T\} \quad (21)$$

O vetor  $w$  representa os scores das folhas,  $q$  é uma função que atribui cada ponto de dados à folha correspondente, e  $T$  é o número de folhas. No XGBoost, a complexidade é definida como:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (22)$$

Evidentemente que há mais de uma forma de definir a complexidade, mas esta tem se mostrado aceitável na prática. A regularização é uma parte que a maioria dos pacotes de árvore trata com menos cuidado, ou mesmo ignorada. Isto porque o tratamento tradicional da aprendizagem em árvore só enfatizava a melhoria da impureza, enquanto o controle da complexidade era deixado para heurísticas. Ao defini-lo formalmente, pode-se ter uma ideia mais apropriada do que está sendo aprendido e construir modelos que apresentem bom desempenho em conjuntos de dados ainda não processados.

#### 2.4.9.2 Score da estrutura

Depois de reformular o modelo de árvore, pode-se escrever o valor objetivo com a árvore  $t$ -ésima como:



$$\begin{aligned}
obj^t &\approx \sum_{i=1}^n [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\
&= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T
\end{aligned} \tag{23}$$

Onde  $I_j = \{i | q(x_i) = j\}$  é o conjunto de índices de pontos de dados atribuídos à folha  $j$ -ésima. Na segunda linha o índice da soma foi alterado porque todos os pontos de dados na mesma folha obtêm a mesma pontuação. Pode-se comprimir ainda mais a expressão definindo  $G_j = \sum_{i \in I_j} g_i$  e  $H_j = \sum_{i \in I_j} h_i$ :

$$obj^{(t)} = \sum_{j=1}^T [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T \tag{24}$$

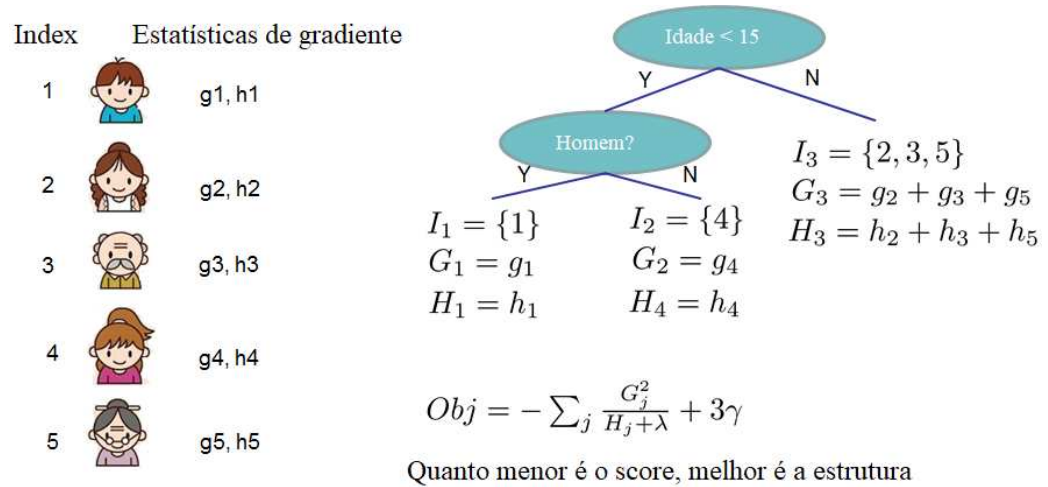
Nesta equação,  $w_j$  são independentes entre si, a forma de  $\sum_{j=1}^T [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2]$  é quadrática e o melhor  $w_j$  para uma dada estrutura  $q(x)$  e a melhor redução da função objetivo que pode-se obter é:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{25}$$

$$obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{26}$$

A última equação mede o quão boa é a estrutura  $q(x)$ . A Figura 12 mostra o cálculo da função objetivo no exemplo do jogo de computador.

Figura 12 – Cálculo de score de estrutura



Fonte: Chen (2016). Traduzido.

Para uma estrutura de árvore dada, calcula-se as estatísticas  $g_i$  e  $h_i$  das folhas a que pertencem, soma-se as estatísticas juntas, e então usa-se a fórmula para calcular o quão boa a árvore é. Esse score é como uma medida de impureza em uma árvore de decisão, exceto que também leva em consideração a complexidade do modelo. Na prática não é viável enumerar todas as possíveis árvores e simplesmente escolher a melhor. Então tenta-se otimizar um nível da árvore de cada vez. Para isso é feito o teste de qual seria o ganho dividindo uma folha em duas novas folhas, conforme mostra a Figura 13.

Figura 13 – Fórmula do ganho na divisão de folhas.

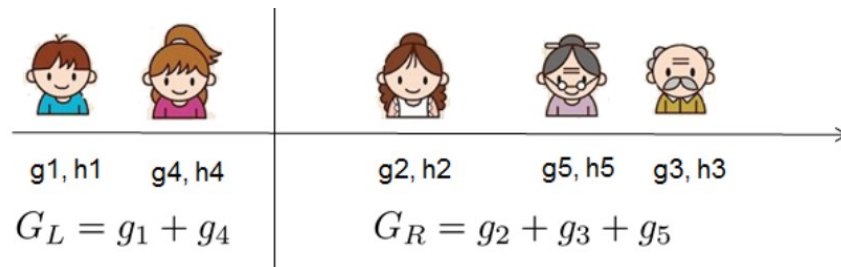
$$\text{Ganho} = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

← O score da folha esquerda
← O score da folha direita
← O score se não for feita a divisão
← O custo de complexidade por adicionar uma folha

Fonte: Chen (2016). Traduzido.

Isso garante que a árvore não seja dividida novamente se não representar um ganho maior do que a árvore sem esse novo ramo. Esta é uma técnica de poda em modelos baseados em árvores que previne que não haja *overfitting*. No exemplo apresentado anteriormente procura-se por uma divisão ideal. Para tal, as instâncias são colocadas de forma ordenada como mostrado figurativamente na Figura 14.

Figura 14 - Exemplo de divisão ideal.



Fonte: Chen (2016). Traduzido.

Uma varredura da esquerda para a direita é suficiente para calcular a pontuação da estrutura de todas as soluções de divisão possíveis, podendo-se encontrar a melhor divisão, de forma eficiente.

#### 2.4.9.3 Vantagens do XGBoost

Desenvolvido por Tianqi Chen em 2016, o *XGBoost*, abreviação de *eXtreme Gradient Boosting*, é uma ferramenta que usa dos conceitos apresentados de *Gradient Boosting* de maneira otimizada para que tenha melhor desempenho computacional. Alguns fatores adicionados à sua construção são:

- I. Paralelização na construção de árvores de decisão usando todos os núcleos dos processadores durante o treinamento;
- II. Computação compartilhada para treino de grandes modelos usando um cluster de máquinas;
- III. *Out-of-Core computation*: Em conjunto de dados muito grandes que não cabem na memória, utiliza-se o disco rígido;
- IV. Otimização do cache das estruturas de dados e algoritmos para fazer melhor uso do *hardware*;
- V. Avaliação da importância das *features*;
- VI. Apresenta um algoritmo para tratar dados faltantes.

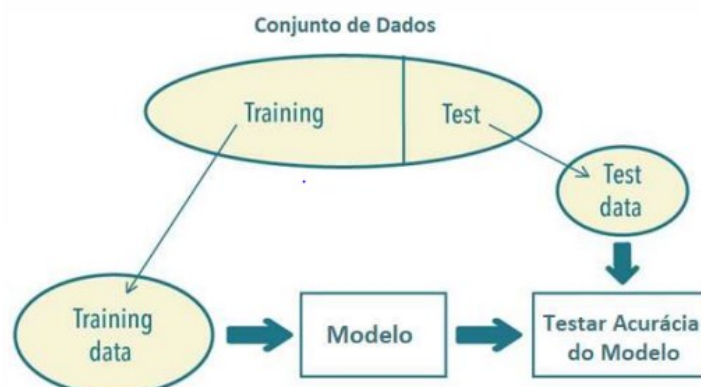
Essas melhorias fizeram com que a implementação do *Gradient Boosting*, que tradicionalmente era lenta por tratar-se da sua natureza sequencial em que as árvores são construídas e adicionadas ao modelo, tornou-se de rápido processamento e portanto, muito popular entre analistas de dados. O *XGBoost* resultou em um dos melhores algoritmos para modelos preditivos, o qual permite extrair todo o potencial das máquinas utilizadas e trazer resultados que foram capazes de vencer competições de aprendizado de máquinas nos últimos anos.

O *XGBoost* apresenta uma série de parâmetros que devem ser otimizados a fim de extrair um melhor resultado do conjunto de dados analisado.

## 2.5 CROSS VALIDATION (VALIDAÇÃO CRUZADA)

O *Cross Validation* serve para avaliar a capacidade de um modelo de generalização, ou seja, busca-se estimar a precisão desse modelo quando usado em um novo conjunto de dados. O treinamento de um modelo de previsão é feito visando diminuir o seu erro. Para que o modelo tenha uma boa capacidade de generalização, é necessário dividir o conjunto de dados em duas partes: o conjunto de dados de treino e conjunto de dados de teste. Conforme ilustrado na Figura 15, o modelo é treinado com o conjunto de dados de treino, no qual tem-se os valores que buscamos prever, para em seguida ter a acurácia testada no conjunto de dados de teste.

Figura 15 - Divisão do conjunto de dados em treino de teste.



Fonte: Fries (2018).

Após serem feitos os testes de acurácia, pode-se gerar um gráfico com os erros de previsão do modelo em relação aos dados de treino e aos dados de teste. Um modelo adequado

de previsão é aquele que consegue minimizar o erro de previsão dos dados de teste, ao mesmo tempo que mantém a complexidade do modelo baixa, a ideia é que se quer estabelecer equilíbrio entre complexidade e precisão (Figura 16).

Figura 16 - Erros de previsão dos dados de teste e treino.

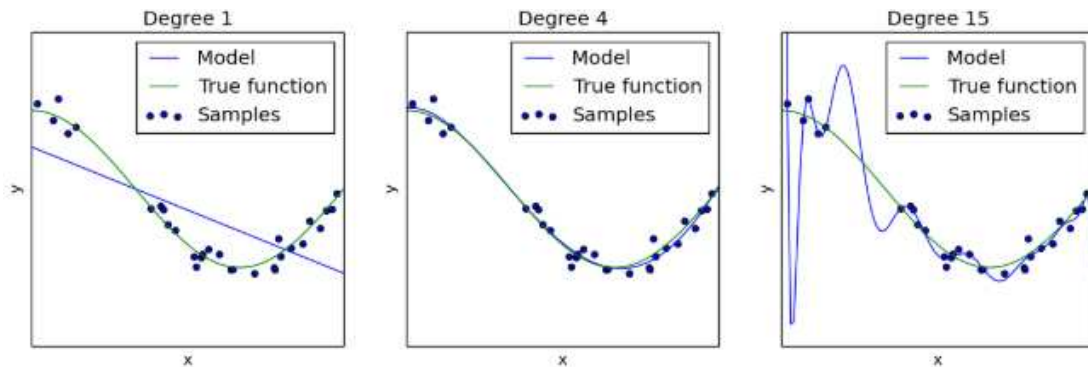


Fonte: Fries (2018).

Como pode-se concluir pela figura, os modelos que sofrem de *underfitting* têm elevados erros em ambos conjuntos de dados. Isso indica que o modelo não consegue gerar uma boa previsão com os dados e algoritmo escolhido. Para melhorar o modelo pode-se aumentar o conjunto de dados de treino, adicionar novas *features*, diminuir a regularização ou, em último caso, usar outro modelo.

Quando o modelo é treinado excessivamente, por diversas épocas, ele se torna bastante assertivo quanto aos dados de treino, porém perde parte da capacidade de generalização. Nesses casos, diz-se que o modelo está com *overfitting*, pois ele aprendeu os dados de treino muito bem, mas quando aplicado em outros conjuntos de dados com especificidades diferentes, apresenta um erro elevado. A solução para esses problemas pode ser simplificar o modelo reduzindo o número de *features*, diminuir o número de épocas no treinamento ou aumentar a quantidade de regularização. A Figura 17 contém exemplos de casos com *underfitting*, *overfitting* e um modelo adequado.

Figura 17 - Exemplos de *underfitting*, um modelo adequado e *overfitting*, respectivamente.



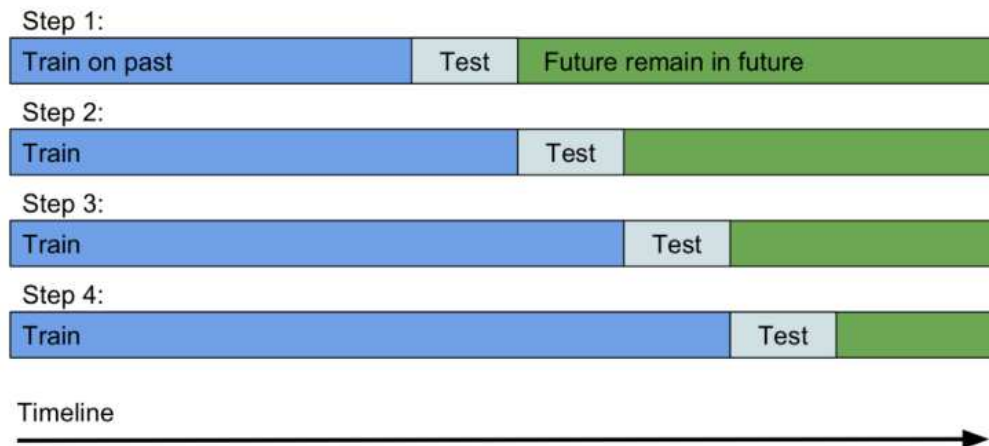
Fonte: Pedregosa et al (2011).

No primeiro gráfico da Figura 17 tem-se um caso clássico de *underfitting* onde uma função de primeiro grau é usada para fazer a previsão, acarretando em um modelo simples, mas que não prevê bem os pontos da amostra. No gráfico do meio consegue-se o menor erro com uma função de 4º grau. No gráfico mais à direita o modelo “decorou” os pontos do conjunto de treino e a função, que é de 15º grau, passa quase que exatamente pelos pontos da amostra. Porém sua capacidade de generalização é baixa e quando comparada a função verdadeira o erro calculado é o pior dentre os três modelos.

É importante notar que os métodos convencionais de validação cruzada onde a base de treino é dividida aleatoriamente durante o treinamento não podem ser usados diretamente com dados de séries temporais. Isto porque eles assumem que não há relação entre as observações, que cada observação é independente. Entretanto, isso não é verdade para os dados de séries temporais, onde a dimensão temporal das observações implica em não poder dividi-los aleatoriamente em grupos. Em vez disso, deve-se dividir os dados e respeitar a ordem temporal em que os valores foram observados.

Esse método de avaliação de modelos de dados históricos é chamado de *Backtesting*. A Figura 18 apresenta um exemplo de *Backtesting*.

Figura 18 – Exemplo de validação cruzada usando *Backtesting*.



Fonte: Osipenko (2018).

## 2.6 MÉTRICAS DE AVALIAÇÃO

As medidas de erros mais comuns para avaliar a precisão e o desempenho de modelos preditivos são o *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), R-Squared ( $R^2$ ) e o *Mean Absolute Percentage Error* (MAPE).

O MAE é uma das principais métricas para avaliação de modelos preditivos e seu cálculo está definido pela Equação 27.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (27)$$

Onde  $y_i$  é o valor de saída esperado e  $\hat{y}_i$  o valor de saída predito pelo modelo. O MAE estima, na média, o erro de previsão do modelo.

O MSE, mostrado pela Equação 28, é uma das métricas mais utilizadas para avaliação de modelos preditivos. Nessa métrica é calculado o erro quadrado médio das previsões, onde quanto maior esse valor, pior é o modelo. Como o MSE eleva os erros ao quadrado, todos os valores são positivos e os erros maiores penalizam mais o modelo. Por esse motivo, faz-se importante um tratamento prévio nos dados para que *outliers* não influenciem de maneira negativa um modelo com bom desempenho.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (28)$$

A métrica RMSE é calculada como a raiz quadrada das médias das diferenças quadradas entre a previsão e o dado real. Em outras palavras, é a raiz quadrada do MSE. A Equação 29 descreve o cálculo do RMSE.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} = \sqrt{MSE} \quad (29)$$

O problema ao utilizar o MSE e RMSE é que é difícil saber se o modelo é suficientemente bom apenas olhando para o seu valor. Um MSE de 15 significa que o modelo é adequado? E 0,9? Para contornar essa situação, pode-se utilizar o coeficiente de determinação  $R^2$ . O  $R^2$  está relacionado ao MSE e apresenta a vantagem de ser livre de escala, não importa se os valores de saída são muito grandes ou muito pequenos, o  $R^2$  sempre estará entre  $-\infty$  e 1. Quando seu valor é negativo, significa que o modelo é pior do que usar a média como previsão. O cálculo do  $R^2$  está descrito na Equação 30.

$$R^2 = 1 - \frac{MSE(\text{modelo})}{MSE(\text{base})} \quad (30)$$

Onde o MSE (base) é calculado substituindo-se o  $\hat{y}_i$ , valor predito, por  $\bar{y}$ , que é a média dos valores observados. Em conclusão,  $R^2$  é a relação entre o quão bom é o modelo quando comparado a modelo que considera a média.

O *Mean absolute percentage error* (MAPE), ou erro percentual absoluto médio, é a média dos erros percentuais absolutos das previsões. O erro é definido como o valor real menos o valor previsto, conforme a Equação 31.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (31)$$

Como são utilizados erros percentuais absolutos, evita-se o problema dos erros positivos e negativos que se anulam mutuamente. Esta medida é fácil de entender porque fornece o erro em termos de porcentagens, e por isso o MAPE tem apelo gerencial e é uma medida comumente utilizada em previsões. Quanto menor a MAPE, melhor a previsão.





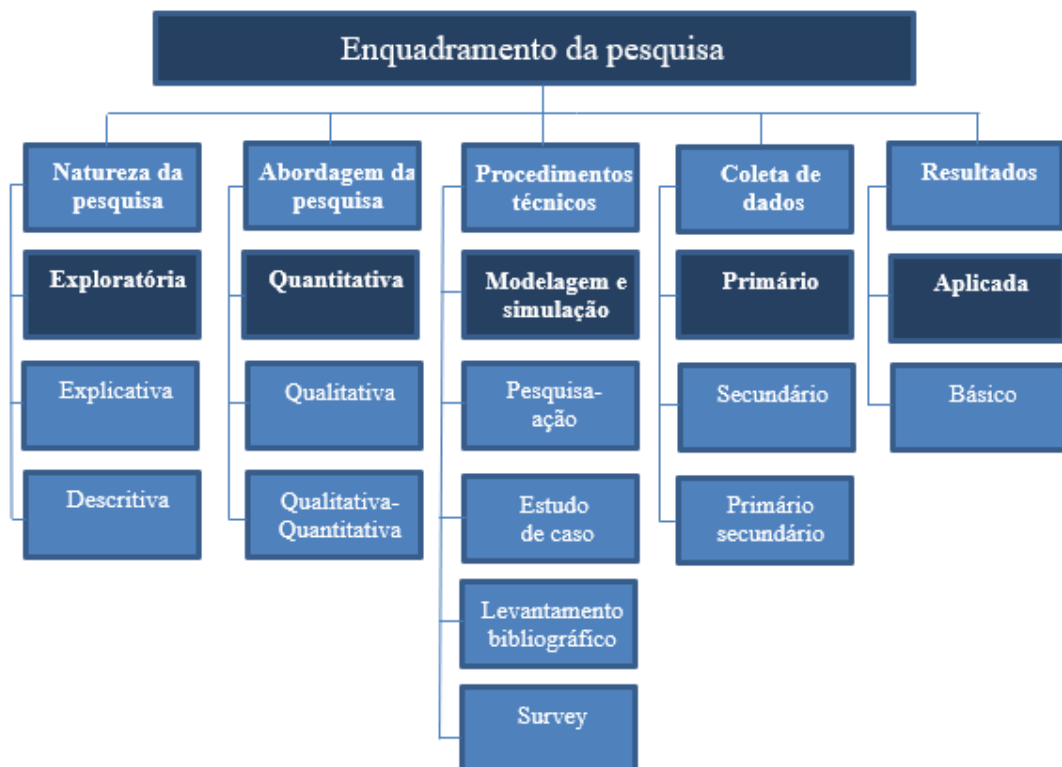
### 3 PROCEDIMENTOS METODOLÓGICOS

Esta seção está dividida em duas partes, a primeira trata de como este trabalho se enquadra nos conceitos de pesquisa, sendo ele classificado conforme os critérios característicos da metodologia científica. Em seguida, o roteiro metodológico é apresentado com uma breve descrição de cada etapa que compõe o estudo.

#### 3.1 ENQUADRAMENTO DA PESQUISA

A natureza da pesquisa é classificada como exploratória, visto que se busca familiaridade com as variáveis mais importantes na predição de vendas do varejo, permitindo a criação de hipóteses acerca do assunto. A classificação conforme os critérios metodológicos está ilustrada na Figura 20.

Figura 20 – Enquadramento da pesquisa.



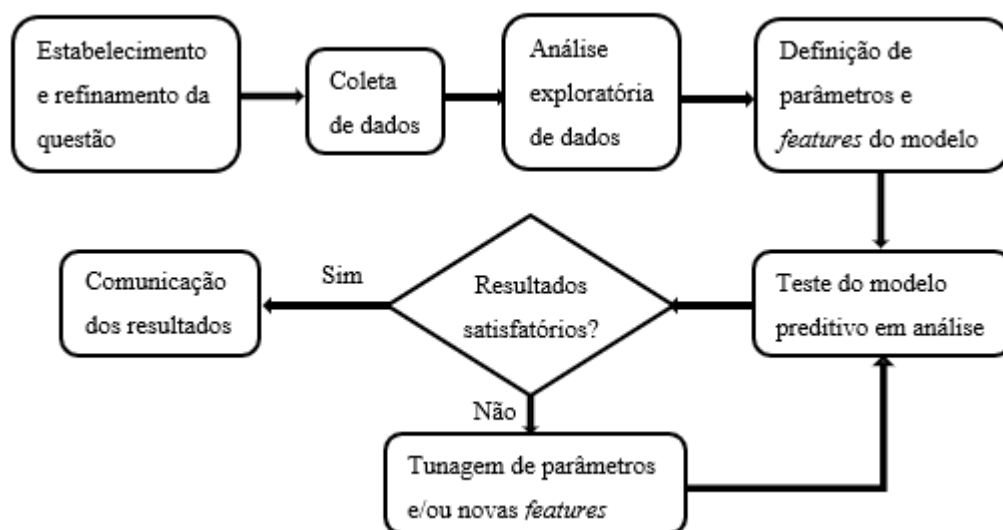
Fonte: Lacerda, Ensslin e Ensslin (2012).

A abordagem é quantitativa, pois utiliza-se de dados estatísticos, métricas de avaliação e análises a partir de modelos matemáticos para se testar o modelo em busca de padrões e melhores previsões. Os procedimentos técnicos enquadram-se em modelagem e simulação, pois é o tipo de pesquisa que compreende o uso de modelos, por meio de técnicas quantitativas para descrever o funcionamento de um sistema ou de parte dele (BERTO; NAKANO, 2000). A coleta de dados foi feita com dados primários, coletados diretamente do banco de dados da empresa objeto de estudo deste trabalho. Os resultados da pesquisa têm objetivo aplicado, visto que se busca um modelo que possa ser utilizado na prática para prever o volume de vendas e auxiliar na tomada de decisão dos gestores.

### 3.2 ROTEIRO METODOLÓGICO

Inicialmente em análise de dados, e também neste trabalho, é preciso estabelecer qual a questão a ser respondida. A escolha é uma etapa fundamental e o tipo de questão influencia a maneira como são executadas as etapas seguintes. A definição do tipo de questão segue os conceitos definidos por Leek (2016) como mostrado na Figura 2, onde o problema deste estudo enquadra-se como uma questão de predição. As etapas que foram executadas no decorrer deste trabalho, a fim de se atingir os objetivos propostos estão ilustradas, no fluxograma da Figura 21.

Figura 21 - Fluxograma do roteiro do trabalho.



Fonte: Elaborado pelo autor (2019).

Definida a questão, é necessário coletar o conjunto de dados que serão usados na análise. Os dados históricos de venda foram então coletados da empresa objeto de estudo desse trabalho por meio de uma consulta no banco de dados utilizando da linguagem SQL (*Structured Query Language*). Foi coletado o histórico de 43 meses de venda de 1221 itens, todos pertencentes ao setor de banho e cozinha da empresa, com informações como loja da venda, estoque, preço e promoções.

Com os dados em mãos, inicia a etapa da análise exploratória de dados. O objetivo dessa análise é examinar os dados previamente à aplicação de qualquer modelo estatístico. Assim, consegue-se ter um entendimento básico dos dados e das relações existentes entre as variáveis. Nessa etapa são visualizadas as características dos dados, como formato do dado, dados faltantes, *outliers*, correlações e medidas estatísticas (maior valor, menor valor, média, mediana, etc.). Para melhor visualizar essas análises, nessa etapa são gerados diversos gráficos e tabelas.

A quarta etapa consiste em criar as *features*, variáveis que servem para ajudar o modelo a aprender melhor o comportamento da série temporal e prever o valor futuro de uma variável de saída. Além disso, a definição dos parâmetros impacta diretamente no desempenho do modelo, sendo a sua parametrização uma etapa crucial.

Com as variáveis definidas, deve-se então testar o modelo preditivo atual baseado em uma métrica de avaliação. Como a performance do modelo é afetada pela forma como ele é parametrizado, é feito um processo de tunagem, ou calibragem, dos parâmetros de forma a experimentar variações diferentes desses parâmetros que levem a um novo modelo o qual tenha capacidade de predição superior ao modelo anterior. Esse processo acontece de forma iterativa e deve ser continuado até que o resultado se mostra satisfatório para o analista, considerando métricas de avaliação pertinentes.

Segundo Peng e Matsui (2016), a última etapa de uma análise de dados é a comunicação dos resultados. É preciso que as conclusões obtidas sejam passadas de maneira clara, completa e que tenha significado para a audiência. O epiciclo proposto pelos autores implica em analisar a forma como os resultados são passados, buscar feedbacks e, se necessário, adequar a apresentação. No presente o trabalho os resultados devem ser apresentados de maneira fácil e clara, a fim de mostrar a execução dos objetivos propostos e o desempenho do modelo. Em razão disso, a ferramenta SHAP (*SHAPley Additive exPlanations*) é usada para ilustrar de maneira visual e intuitiva o grau de importância das variáveis no modelo de predição (LUNDBERG; LEE, 2017).

## 4 CONTEXTUALIZAÇÃO DO PROBLEMA

Este capítulo apresenta a empresa, o objeto de estudo deste trabalho bem como as características dos itens escolhidos para análise e as condições do mercado nas quais a empresa está inserida.

### 4.1 A EMPRESA

O grupo foi fundado em 1958 com a abertura de uma serraria em Urubici - SC. De acordo com o site da empresa, no final da década de 1960, foi criada uma madeireira, estabelecida em São José - SC. Após alguns anos, em 1967, decidiu-se reestruturar a estratégia da organização e foi inaugurada a primeira loja com característica de *Varejo Centerlar*, sendo esta a filial de Campinas, em São José - SC. Ao final da década de 1970, ocorreu um momento de expansão na companhia, com a abertura de novas unidades.

Atualmente a empresa conta com 19 filiais distribuídas por três estados do país (Figura 22), sendo estes: Santa Catarina, Paraná e Rio Grande do Sul. Possui hoje três Centro de Distribuição, um em cada estado, onde são armazenados seus itens com o intuito de abastecer as filiais regionais. A organização conta com um mix de aproximadamente 45 mil itens e tem hoje o maior estoque a pronta-entrega do Sul do país. Além disso, a companhia emprega hoje cerca de 4 mil empregados diretos.

Em 2018, ainda de acordo com o site da organização, foi eleita pelo segundo ano consecutivo como a quinta maior empresa de material de construção do Brasil pelo ranking Anamaco. Conta hoje com mais de 90 fornecedores mundiais e apresenta uma linha completa de marcas próprias exclusivas. Por fim, a empresa tem suas atividades operacionais voltadas para atender ao seu propósito final: colaborar com as pessoas para transformar a casa em um lar.

Figura 22 - Localização das lojas e centros de distribuição.



Fonte: Elaborado pelo autor (2019).

#### 4.2 ITENS ESCOLHIDOS PARA A ANÁLISE

Dos dados coletados referentes a 1.221 itens, foram escolhidos quatro itens que apresentaram séries temporais de comportamento diferenciado. Os itens que foram selecionados (mostrados na Figura 23) para avaliar a capacidade de previsão do modelo, sob condições diversas, consistiram em:

- I. Parafuso para fixação de bacias e bidês Deca 10mm com duas peças cromado.
- II. Tubo de ligação plástico Tigre
- III. Kit instalação para aquecedor a gás 1/2" 40cm Jackwal
- IV. Assento sanitário almofadado convencional/oval branco Astra

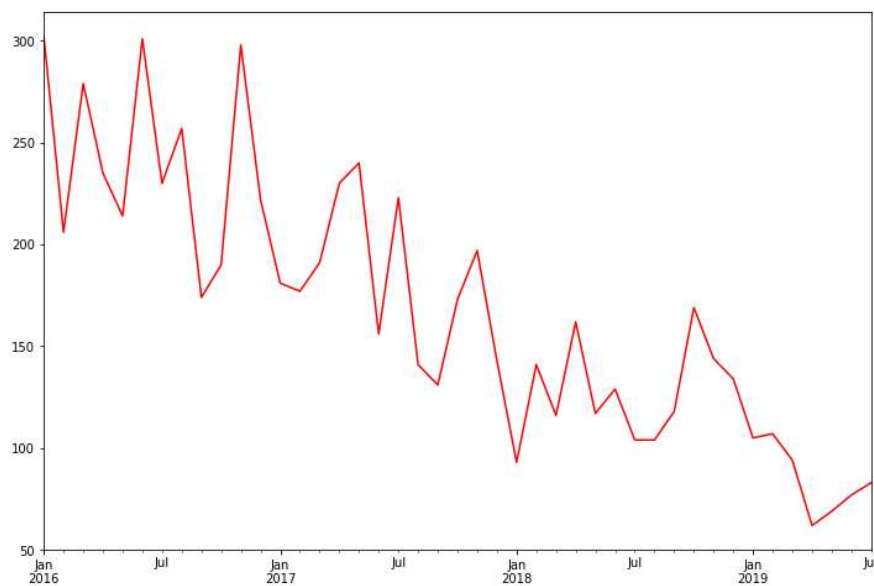
Figura 23 - Itens selecionados para análise.



Fonte: Empresa.

Os dados históricos de vendas desses itens foram levantados para o período de janeiro/2016 a julho/2019 compreendendo 43 meses consecutivos de todas as lojas da empresa. O item I, parafuso para fixação, apresenta uma tendência decrescente nas vendas no passado recente, conforme pode ser constatada na Figura 24.

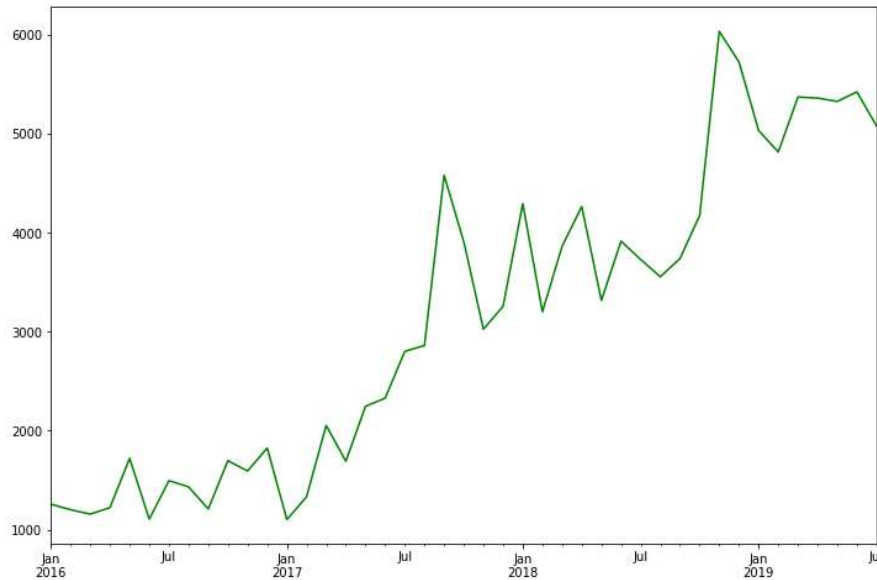
Figura 24 - Histórico mensal de vendas do item I (parafuso de fixação).



Fonte: Elaborado pelo autor (2019).

O item II, tubo de ligação, tem um comportamento de venda com tendência crescente no período especificado de 43 meses. A Figura 25 mostra o histórico de vendas desse item.

Figura 25 - Histórico mensal de vendas do item II (tubo de ligação).

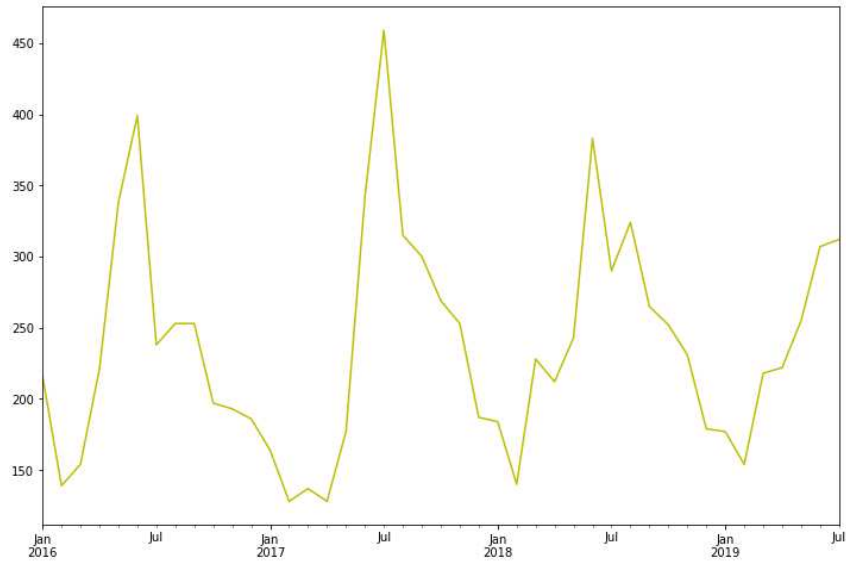


Fonte: Elaborado pelo autor (2019).

O kit de instalação para aquecedor a gás, item III, apresenta visivelmente uma característica sazonal no seu histórico de vendas conforme mostrado na Figura 26.



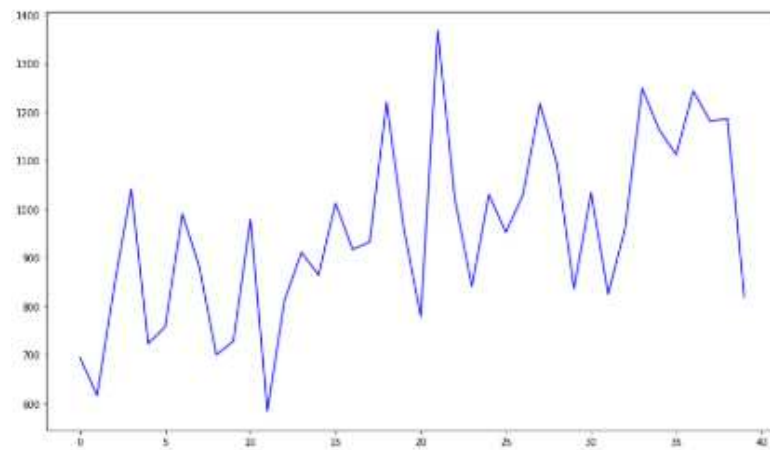
Figura 26 - Histórico mensal de vendas do item III (kit de instalação para aquecedor a gás).



Fonte: Elaborado pelo autor (2019).

O assento sanitário (item IV), cuja série temporal do histórico de vendas é mostrada na Figura 27, apresenta comportamento diferenciado dos demais no sentido de que não apresenta uma forte tendência nem sazonalidade aparente.

Figura 27 - Histórico mensal de vendas do item IV (assento sanitário).



Fonte: Elaborado pelo autor (2019).

## 5 DESENVOLVIMENTO

O presente capítulo trata do processo de desenvolvimento dos modelos de previsão, com foco nas macroetapas II e III do epícciclo da análise de dados, exploração de dados e construção do modelo formal.

### 5.1 ANÁLISE EXPLORATÓRIA

A etapa da análise exploratória serve para familiarizar-se com os dados, organizá-los e sintetizá-los de forma a começar a criar as primeiras hipóteses.

#### 5.1.1 Análise do conjunto inicial de dados

O conjunto de dados inicial coletado do banco de dados da empresa objeto de estudo refere-se ao histórico mensal de 1.221 itens do setor de banho e cozinha, no período de 01/01/2016 a 31/07/2019 (total de 43 meses). As informações disponíveis nesse banco de dados estão mostradas exemplarmente na Figura 28.

Figura 28 - Conjunto de dados inicial.

```
df.head()
```

	CODITEM	MES	CODFIL	VENDA	RECEITA	PREÇO_MEDIO	DIAS	PARTICIPAÇÃO	RUPTURA
0	1984	2016-01-01	1	2.0	139.8	69.9	31.0	0.065385	0.0
1	1984	2016-01-01	3	NaN	NaN	NaN	31.0	0.000000	0.0
2	1984	2016-01-01	8	1.0	59.9	64.9	31.0	0.130769	0.0
3	1984	2016-01-01	10	3.0	239.7	79.9	31.0	0.103846	0.0
4	1984	2016-01-01	11	NaN	NaN	NaN	31.0	0.276923	0.0

Fonte: Elaborado pelo autor (2019).

Cada linha contém as informações de um determinado item, identificado pelo seu código, em uma filial e em um determinado mês. Tem-se também quantas unidades foram vendidas e a receita gerada. A coluna “RUPTURA” indica em quantos % dos dias do mês o item estava em ruptura naquela loja. A variável “PARTICIPAÇÃO”, por sua vez, indica quanto a receita da loja tem de participação na receita anual gerada por determinado item. Ou seja, as

lojas maiores tendem a ter uma maior participação nas vendas e por isso a ruptura nessas lojas tem maior impacto que a ruptura em lojas menores.

A biblioteca Pandas do Python tem ferramentas estatísticas que dão informações acerca dos dados analisados. A Figura 29 mostra o resultado do comando “info()” que retorna o número de linhas do conjunto de dados, número de linhas de cada coluna e o formato no qual o dado está configurado. Esses dados evidenciam a presença de dados faltantes e/ou dados em formatos indevidos ou indesejados. Do resultado dessa análise, destaca-se o número de linhas não-nulas da coluna “VENDA”, que é de 556.159 linhas. Esse valor indica que em apenas 53,41% dos registros coletados houve venda do item naquele mês na filial.

Figura 29 - Comando “info()” aplicado no conjunto de dados inicial.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1041362 entries, 0 to 1041361
Data columns (total 9 columns):
CODITEM      1041362 non-null int64
MES          1041362 non-null datetime64[ns]
CODFIL      1041362 non-null int64
VENDA       556159 non-null float64
RECEITA     556159 non-null float64
PREÇO_MEDIO 556159 non-null float64
DIAS        1027742 non-null float64
PARTICIPAÇÃO 1041362 non-null float64
RUPTURA     1041362 non-null float64
dtypes: datetime64[ns](1), float64(6), int64(2)
memory usage: 71.5 MB
```

Fonte: Elaborado pelo autor (2019).

O comando “.describe()” da biblioteca Pandas traz informações estatísticas de cada coluna quantitativa do conjunto de dados. Além da contagem de linhas, fornece também a média, o desvio padrão, os quartis e o valor mínimo e máximo. A Figura 30 mostra as estatísticas da base analisada. A média da variável “RUPTURA” indica que em 13,05% dos dias houve falta de estoque de algum dos itens considerados naquela filial.

Figura 30 - Dados estatísticos do conjunto de dados inicial.

	VENDA	PREÇO_MEDIO	DIAS	PARTICIPAÇÃO	RUPTURA
<b>count</b>	556159.00000	556159.00000	1027742.00000	1041362.00000	1041362.00000
<b>mean</b>	6.76624	171.09349	26.79312	0.07046	0.13053
<b>std</b>	18.67642	258.34452	7.63417	0.08747	0.26774
<b>min</b>	0.00000	0.10000	1.00000	0.00000	0.00000
<b>25%</b>	1.00000	39.90000	28.00000	0.01591	0.00000
<b>50%</b>	2.00000	99.90000	30.00000	0.04817	0.00000
<b>75%</b>	6.00000	209.90000	31.00000	0.09348	0.06452
<b>max</b>	993.00000	6999.00000	31.00000	1.00000	1.00000

Fonte: Elaborado pelo autor (2019).

Como a questão a ser respondida é a previsão de venda dos itens por mês na rede, agrupou-se os dados consolidando as informações que antes eram por filial, de forma a ter apenas uma linha por mês de cada item. O conjunto de dados agrupado está mostrado exemplarmente na Figura 31.

Figura 31 - Parte do conjunto de dados agrupado.

	item	mes	Venda
<b>0</b>	44561	jan/2016	302
<b>1</b>	74315	jan/2016	216
<b>2</b>	86882	jan/2016	1255
<b>3</b>	129794	jan/2016	694
<b>4</b>	44561	fev/2016	203
<b>5</b>	74315	fev/2016	139
<b>6</b>	86882	fev/2016	1201
<b>7</b>	129794	fev/2016	616

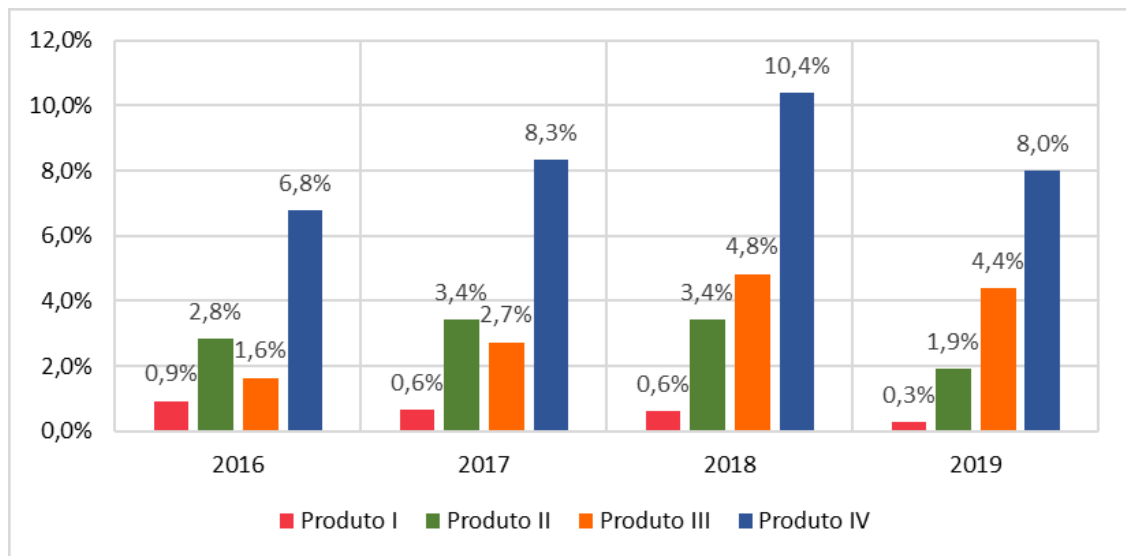
Fonte: Elaborado pelo autor (2019).

Com isso o novo conjunto de dados agrupado traz a soma de toda a venda do item naquele mês, podendo essa venda ser comparada a outros valores como exercício de comparação. Para se ter noção da importância dos itens escolhidos em termos de receita, foi calculada a participação das vendas de cada item dentro de sua família de itens, por ano.

A Figura 32 mostra que o item IV apresenta alta participação nas vendas da família de

itens “Assentos Sanitários”, assim como o item III que é da mesma família. O item I representa menos de 1% das vendas da família “Assentos sanitários”. O item II pertence à família de “Aquecedores e exaustão” e sua participação nas vendas vem crescendo ao longo dos primeiros três anos do período de análise. Vale destacar que os dados referentes ao ano de 2019 incluem vendas até o mês de julho.

Figura 32 - Participação dos itens dentro da sua família de itens.



Fonte: Elaborado pelo autor (2019).

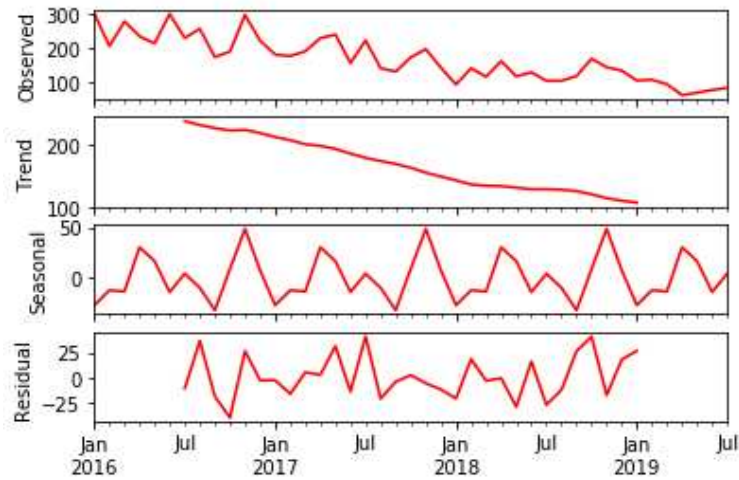
### 5.1.2 Decomposição das séries temporais

A biblioteca *Statsmodel* é um módulo do *Python* que fornece classes e funções para estimar diversos modelos estatísticos. A decomposição de séries temporais está disponível como recurso nessa biblioteca e pode-se aplicá-la aos históricos de vendas dos itens analisados. A decomposição baseia-se no modelo aditivo onde a série é representada pela função  $Y(t)$ , calculada como a soma da componente de tendência  $T(t)$ , de sazonalidade  $S(t)$  e a componente residual  $e(t)$ .

$$Y(t) = T + S(t) + e(t) \quad (32)$$

A Figura 33 mostra a decomposição da série temporal do item I.

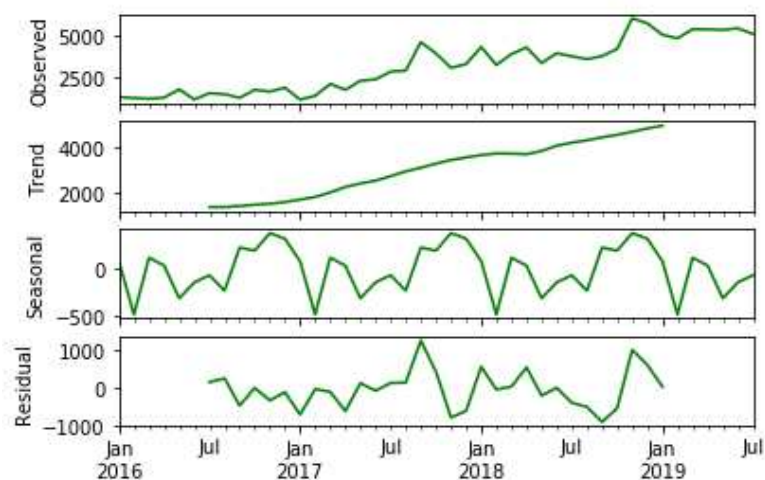
Figura 33 - Decomposição do histórico mensal de vendas do item I.



Fonte: Elaborado pelo autor (2019).

A decomposição dessa série temporal evidencia que o histórico de vendas apresenta forte tendência decrescente. A componente sazonal identificada não é tão impactante quanto à tendência. Quanto à decomposição da série temporal referente ao item II, mostrada na Figura 34, indica alta tendência crescente, baixa sazonalidade e grande nível de resíduo, característica de uma série temporal menos regular.

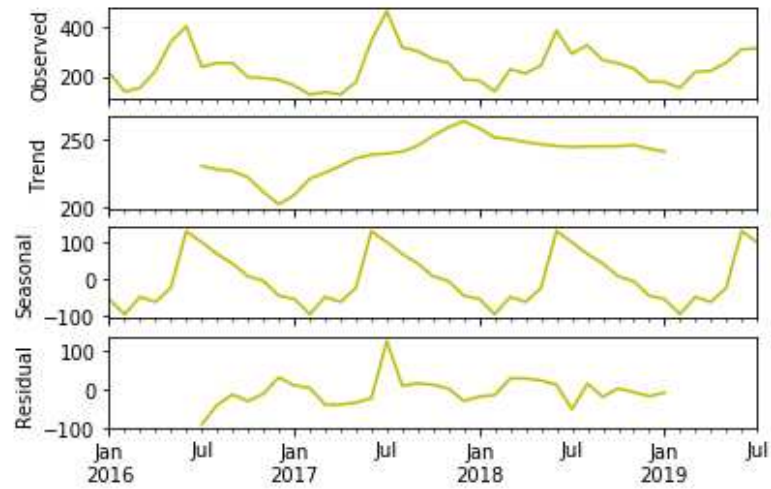
Figura 34 - Decomposição do histórico mensal de vendas do item II.



Fonte: Elaborado pelo autor (2019).

O item III tem componente sazonal bem presente, assim como a componente residual. A tendência não se mostra tão relevante para essa série temporal.

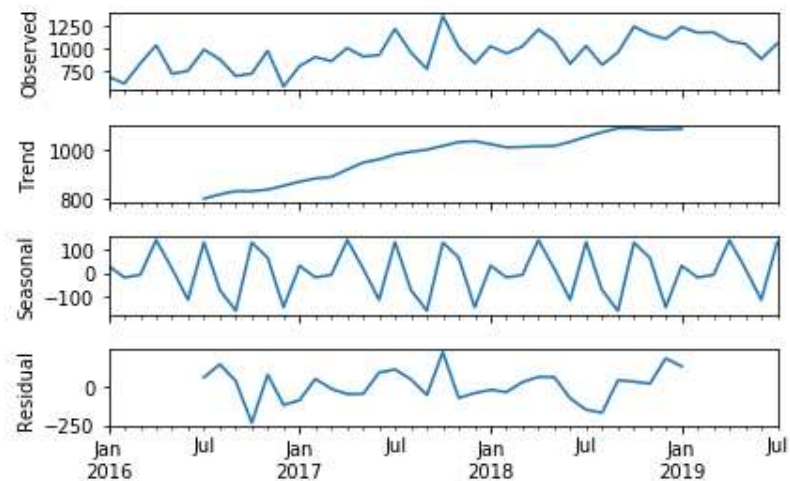
Figura 35 - Decomposição do histórico mensal de vendas do item III.



Fonte: Elaborado pelo autor (2019).

No histórico mensal de vendas do item IV, as três componentes de séries temporais são relevantes e fundamentais na soma que forma a série original. A Figura 36 mostra a decomposição do item IV.

Figura 36 - Decomposição do histórico mensal de vendas do item IV.



Fonte: Elaborado pelo autor (2019).

## 5.2 GERAÇÃO DE *FEATURES*

Com o intuito de aperfeiçoar o modelo de previsão do *XGBoost*, criou-se novas *features* a partir das informações coletadas do banco de dados da empresa. As *features* servem para adicionar informações às séries temporais de forma que o modelo aprenda como essas informações influenciam na variável que se deseja prever.

O processo de geração de *features* em séries temporais necessita de atenção com relação a “vazamento” de dados. O vazamento ocorre quando os dados que você está usando para treinar um algoritmo de aprendizado têm as informações que você está tentando prever.

### 5.2.1 Ruptura Ponderada

As informações de ruptura, ou seja, de falta de estoque do item na loja, foram agrupadas de forma ponderada. Inicialmente foi calculada a porcentagem de ruptura dos itens mensalmente por filial, dividindo o número de dias em que não se tinha o item na loja pelo número de dias do mês. Em seguida, agrupou-se as vendas das filiais de maneira ponderada para expressar a disponibilidade do item em determinado mês considerando todas as lojas. A ponderação levou em conta a participação na receita daquele item no ano por filial. Isso quer dizer que a ruptura em uma filial que teve maior receita tem maior peso sobre a ruptura em uma filial cuja participação na receita foi menor.

A *feature* “Ruptura” é então a multiplicação da porcentagem de ruptura da filial naquele mês com a porcentagem de participação da receita da loja na receita total naquele ano.

### 5.2.2 Variação do preço médio

A *feature* “Var Preço” tem o objetivo de trazer a informação sobre a influência do preço do item nas vendas. Como a base de treino contém informações de mais de 1.000 itens, o valor do preço do item como *feature* não serve ao modelo, visto que a faixa de preço dos itens analisados varia bastante e os valores não podem ser comparados entre si. No entanto, pode-se comparar o preço do item no mês em relação à sua série histórica.

Dessa forma, a variável “Var preço” é calculada como a variação percentual do preço médio do mês em questão em relação ao mês passado. O preço médio é calculado como a média



de preço das unidades vendidas em determinado mês considerando todas as lojas. “Var Preço” é positivo quando há aumento de preço e negativo quando o preço diminui.

$$Var\ Preço = \frac{Preço\ médio_t - Preço\ médio_{t-1}}{Preço\ médio_t} \quad (33)$$

### 5.2.3 Vendas em tempos passados

As *features* de venda defasadas em valores passados são essenciais para a previsão de séries temporais. A *feature* consiste nos valores de venda do item defasados em “n” períodos de tempo, conhecidos como “lags”. O período de tempo adotado no modelo corresponde aos 13 últimos meses de vendas. O recurso permite ao modelo aprender melhor o comportamento da série temporal de sazonalidade e tendência.

### 5.2.4 Média por família

A *feature* “MédiaFamília\_1” é calculada a partir da média de venda dos itens da família do item no mês anterior. A coluna “Família” não poderia ser usada no modelo por tratar-se de uma variável no formato de texto, então se usa esse recurso de trazer a média da variável que se deseja prever a fim de incluir a informação da família do item no modelo.

### 5.2.5 Variação em relação ao ano passado

A *feature* “Var LY” cria uma previsão de venda partindo do princípio que a variação ocorrida entre o mês passado e treze meses atrás irá se repetir no mês atual. A Equação 34 apresenta o cálculo dessa *feature*.

$$Var_{LY} = \frac{(Venda\ t-1)}{(Venda\ t-13)} * (Venda\ t - 12) \quad (34)$$

## 5.3 TUNAGEM/OTIMIZAÇÃO DE HIPERPARÂMETROS

A tunagem ou otimização de hiperparâmetros consiste em escolher um conjunto de hiperparâmetros ótimos para o algoritmo de aprendizado. Com as *features* criadas, pode-se

calibrar os parâmetros a fim de alcançar resultados mais assertivos. A biblioteca *sklearn* conta com uma função chamada “*GridsearchCV*” que permite testar o modelo dentro de um conjunto de hiperparâmetros definidos pelo analista. A Tabela 3 apresenta os hiperparâmetros utilizados na tunagem e o valor ótimo encontrado para os mesmos.

Tabela 3 – Tunagem de hiperparâmetros.

Hiperparâmetro	Descrição	Valores testados	Melhor valor
colsample_bytree	Fração das colunas que serão utilizadas em cada árvore.	[0.5, 0.6, 0.7, 0.8, 0.9]	0.8
learning_rate	Diminui o peso em cada etapa para evitar <i>overfitting</i> .	[0.01, 0.1, 0.3]	0.1
max_depth	Profundidade máxima da árvore. Árvores profundas correm o risco de apresentar <i>overfitting</i> .	[3, 5, 7, 9, 11]	9
min_child_weight	Indica a soma mínima dos pesos de todas as observações de uma mesma folha.	[2, 3, 5, 7, 9]	3
n_estimators	Número de árvores usadas no modelo.	[100,200]	100
subsample	Indica a fração de observações usadas como amostra para cada árvore.	[0.6, 0.8, 1]	1

Fonte: Elaborado pelo autor (2019).

## 5.4 MODELOS ESTATÍSTICOS

Os modelos estatísticos utilizados para fins de comparação de resultados foram Holt-Winters e SARIMA. Esses modelos de previsão diferem do XGBoost por usarem apenas as séries históricas de vendas. A base de dados apresenta duas colunas: data (mês) e vendas.

### 5.4.1 Holt-Winters

A biblioteca *Statsmodel* fornece o algoritmo que treina um modelo de previsão usando o método Holt-Winters para em seguida utilizá-lo para prever valores no horizonte de tempo desejado. O modelo é então treinado usando os 41 primeiros meses de venda e depois a previsão é feita para os próximos 2 meses. Os parâmetros que precisam ser definidos são: o tipo da

componente de tendência, o tipo da componente de sazonalidade e o número de períodos que um ciclo completo de sazonalidade. Os tipos de componentes podem ser aditivos ou multiplicativos e a escolha do tipo utilizado foi a que apresentou menores erros de previsão. O número de períodos de sazonalidade utilizado foi 12 para os quatro produtos. Na Tabela 4 estão apresentados os parâmetros do modelo Holt-Winters para cada um dos quatro itens.

Tabela 4 – Parâmetros Holt-Winters.

Itens	Tendência	Sazonalidade	Ciclo sazonal (meses)
Item I	Aditivo	Aditivo	12
Item II	Aditivo	Aditivo	12
Item III	Aditivo	Aditivo	12
Item IV	Aditivo	Multiplicativo	12

Fonte: Elaborado pelo autor (2019).

#### 5.4.2 SARIMA

O método de previsão SARIMA também está contido na biblioteca *Statsmodel* e, como no uso do Holt-Winters, a base é dividida entre treino e teste, sendo os 41 primeiros meses usados para treinar o modelo que então faz a previsão para os próximos 2 meses.

Para encontrar os parâmetros  $(p, d, q)(P, D, Q)_m$  que definem o modelo SARIMA, utilizou-se a biblioteca *pmdarima* que contém a função chamada *auto\_arima*. O *auto\_arima* auxilia o analista de dados ao fornecer os parâmetros ótimos para a série temporal de acordo com o menor AIC (*Aikaike information criterion*), critério que avalia modelos estatísticos de acordo com o erro e complexidade do modelo. Os parâmetros selecionados para os modelos SARIMA de cada um dos quatro itens estão apresentados na Tabela 5.

Tabela 5 – Parâmetros SARIMA.

Itens	$(p, d, q)$	$(P, D, Q)_m$
Item I	(0,1,1)	$(0,0,0)_0$
Item II	(1,1,1)	$(0,0,1)_{12}$
Item III	(1,0,0)	$(2,0,0)_{12}$
Item IV	(0,1,1)	$(0,0,1)_{12}$

Fonte: Elaborado pelo autor (2019).

## 6 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

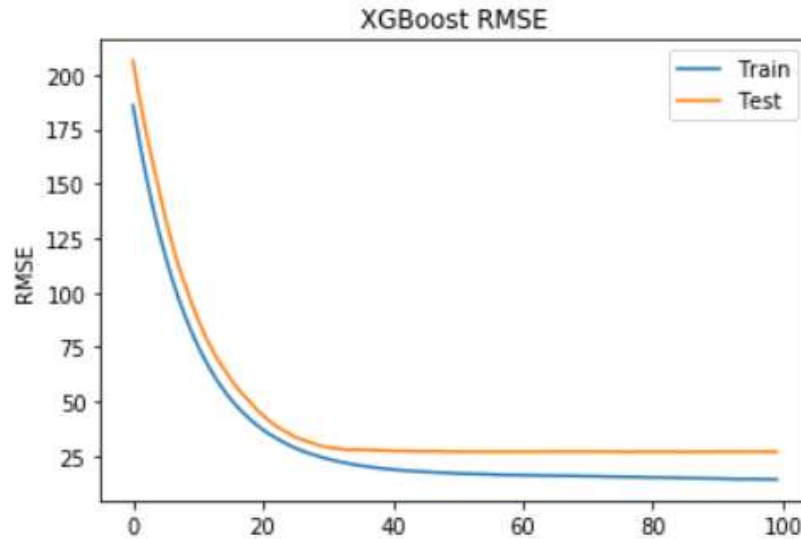
Neste capítulo são apresentados os resultados obtidos com os modelos de previsão para cada um dos quatro itens analisados. As métricas utilizadas para avaliar a precisão dos modelos foram RMSE e MAPE. A importâncias das variáveis estão apresentadas no formato do XGBoost e do SHAP.

### 6.1 VERIFICAÇÃO DE *OVERFITTING*

O modelo final após as otimizações de parâmetros e geração de features foi treinado com os 41 primeiros meses da base de treino. Durante o treino pôde-se calcular o erro do modelo conforme ele vai aprendendo com os dados. Para verificar se o modelo manteve sua capacidade de generalização e não houve *overfitting*, pode-se plotar o erro obtido nas bases de treino e teste ao longo do treino.

No eixo X está representado o número de árvores incluídas no modelo. Conforme o modelo vai aprendendo com os erros de cada árvore, uma árvore seguinte é adicionada de forma a corrigir os erros anteriores. No eixo Y está o erro entre o valor previsto do modelo e o valor real, metrificado pelo RMSE. Em azul têm-se os erros da base de treino e em laranja os da base de teste. É importante que a diferença entre os erros dessa base de treino, que foi a utilizada pelo modelo no seu treinamento, e a base de teste, com dados nunca antes vistos pelo modelo, seja a menor possível. Analisando a Figura 37, conclui-se que o modelo não sofre de *overfitting* e é possível notar que após a 30ª árvore adicionada, não houve grandes ganhos em termos de redução de erro de previsão.

Figura 37 - Erros de teste e treino



Fonte: Elaborado pelo autor (2019).

## 6.2 ANÁLISE GERAL DOS MODELOS

O modelo de previsão teve como objetivo prever o volume de vendas dos meses de junho de 2019 e julho 2019. Os resultados obtidos foram coletados e em seguida calculados os erros MAPE e RMSE. Na Tabela 6 são apresentados os RMSEs dos três modelos para os itens analisados.

Tabela 6 – RMSE dos métodos por item.

Item	Holt-Winters	SARIMA	XGBoost
Item I	11,6	6,6	6,5
Item II	301,0	181,1	191,9
Item III	86,7	107,2	24,6
Item IV	151,0	92,7	169,3

Fonte: Elaborado pelo autor (2019)

Os valores de RMSE dos modelos apontam o SARIMA como o modelo que obteve melhor desempenho na predição dos itens II e IV. O Xgboost conseguiu performar melhor nos itens I e III, enquanto o método Holt-Winters não apresentou os melhores resultados para nenhum dos itens analisados. O RMSE é uma métrica caracterizada por punir grandes erros e

seu valor é proporcional a magnitude dos dados, tornando difícil a comparação entre os itens, visto que os volumes de vendas reais variam de 77 a 5.425 unidades.

O MAPE (erro absoluto percentual médio) apresenta o erro do modelo em termos percentuais em relação ao volume real de vendas. Dessa forma fica evidente a precisão do modelo preditivo. A Tabela 6 apresenta os MAPEs dos três modelos de previsão para os itens analisados.

Tabela 7 – MAPE dos métodos por item.

Item	Holt-Winters	SARIMA	XGBoost
Item I	21,0%	7,1%	5,3%
Item II	5,5%	3,4%	2,7%
Item III	16,7%	32,9%	7,5%
Item IV	16,7%	7,6%	14,8%
Média	15,0%	12,8%	7,58%

Fonte: Elaborado pelo autor (2019).

Os resultados obtidos pelo Xgboost são, na média, melhores que os dos demais modelos, com um erro médio de 7,58% em relação ao volume de vendas observado.

O modelo SARIMA obteve o maior MAPE na previsão de venda do item III, mas teve bom desempenho comparado aos outros modelos na previsão dos itens I, II e obteve o melhor resultado na previsão do item IV. Por outro lado, o modelo Holt-Winters mostrou-se pouco preciso em prever a venda dos itens analisados.

### 6.3 IMPORTÂNCIA DAS FEATURES

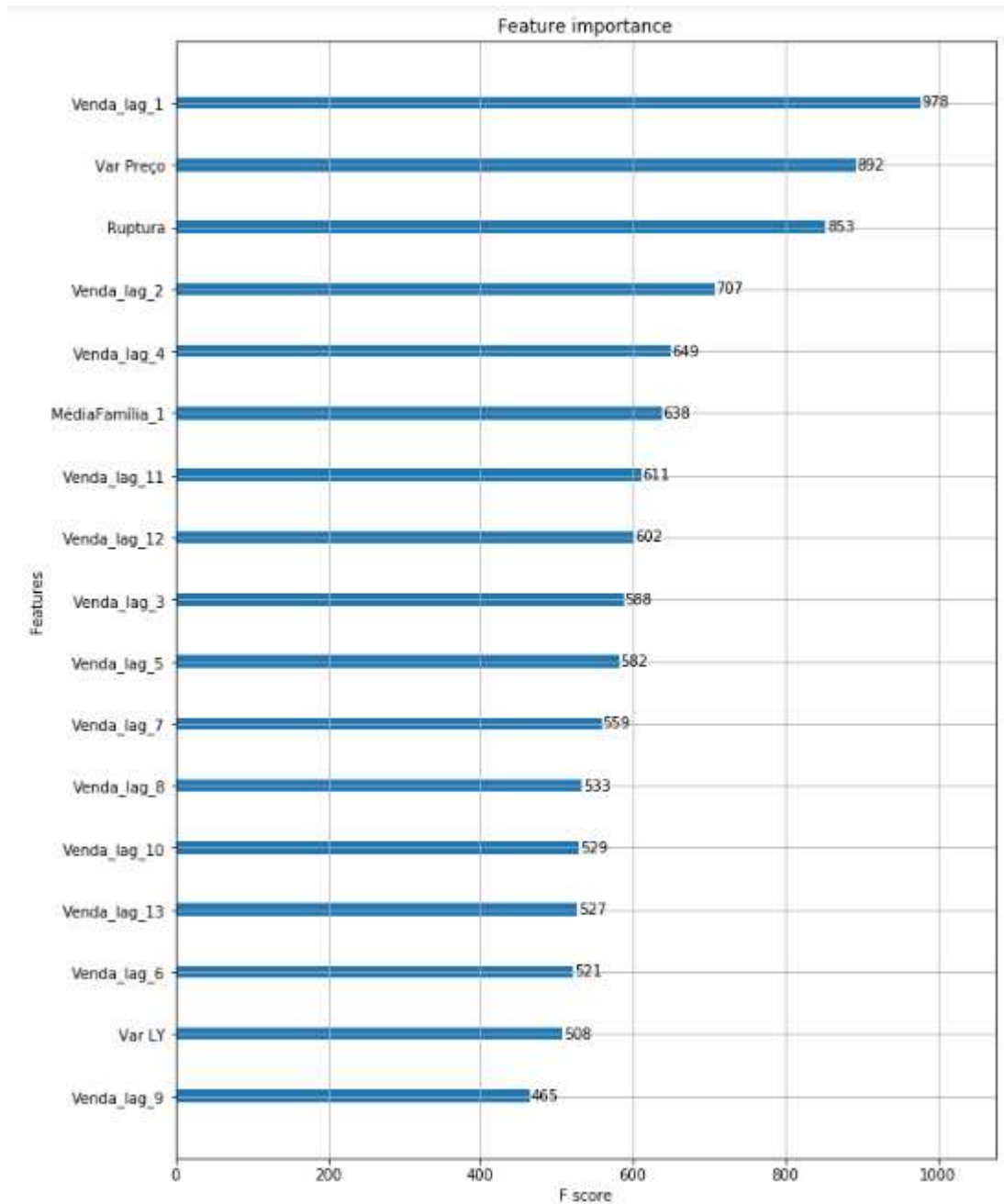
Os resultados obtidos pelo XGBoost são calculados a partir das *features*. Saber quais são as que tiveram mais impacto na previsão ajuda o analista de dados a entender que tipo de informação gerada o modelo de fato está usando na previsão e quais tiveram pouca ou nenhuma importância.

A análise da importância das *features* auxilia também a entender o fenômeno que está sendo analisado. Pode-se retirar *insights* valiosos com essas informações de *features* que não aparentam ter relação óbvia com a variável prevista, mas que o modelo indica que tem forte relação.

A biblioteca do XGBoost apresenta uma opção que gera o gráfico da importância de

cada *feature* para o modelo. A Figura 38 mostra quais foram as importâncias para as *features* utilizadas no XGBoost.

Figura 38 – Importância das *features* no modelo de previsão.



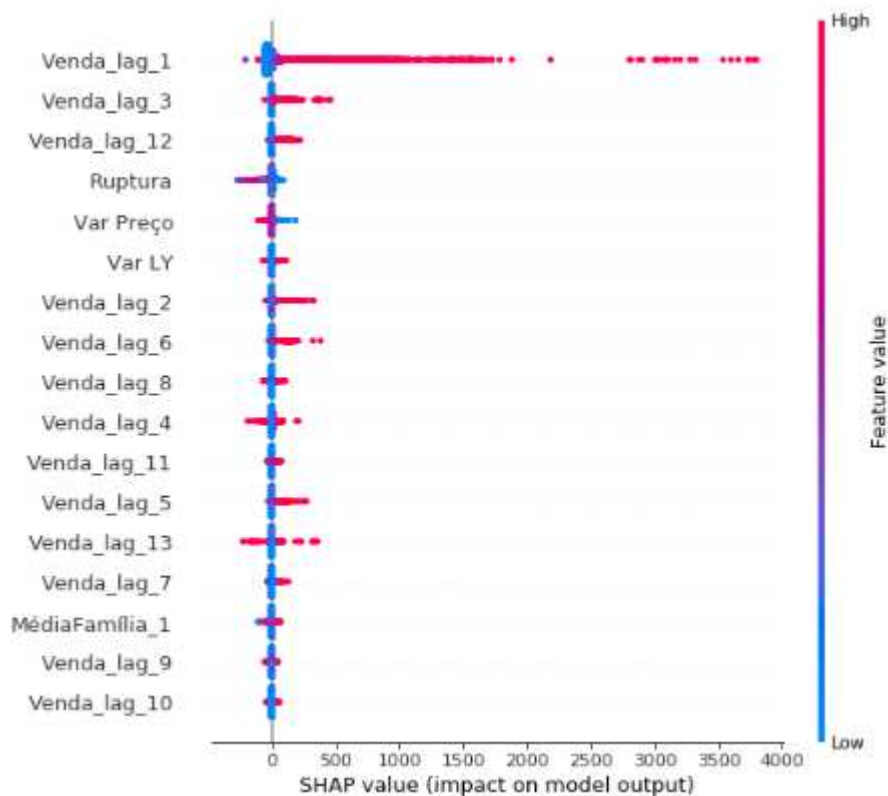
Fonte: Elaborado pelo autor (2019).

A variável mais importante identificada foi “Venda\_lag\_1”. O fato dessa feature, que representa a venda do mês anterior, ter o maior impacto no modelo já era esperado, visto que a

demanda tende a não se alterar tanto de um mês para o outro seguinte. A *feature* “Var Preço” obteve um alto *score* de importância devido à influência que o preço do item tem nas vendas. Uma variação negativa no preço tende a gerar mais vendas enquanto uma variação positiva tende a fazer com que as vendas diminuam. Em terceiro no ranking de importância das *features* está a “Ruptura”. A ruptura de um item é um fenômeno que tende a diminuir as suas vendas.

A ferramenta SHAP permite uma forma alternativa de visualização da importância das *features* do modelo. As *features* estão ordenadas em grau de importância de cima para baixo no eixo Y e o impacto das mesmas no modelo é mostrado no eixo X do gráfico. A cor indica o valor da *feature*, se é alto ou baixo (Figura 39).

Figura 39 Grau de importância das *features* com a ferramenta SHAP.



Fonte: Elaborado pelo autor (2019).

A feature “Venda\_lag\_1” novamente aparece como a mais importante do modelo, seguida da “Venda\_lag\_3” e “Venda\_lag\_12”. A venda defasada em 12 meses é importante principalmente em itens com sazonalidade e sem tendência, nesses casos a venda do mesmo mês no ano anterior pode aproximar-se mais do que a venda de meses recentes.

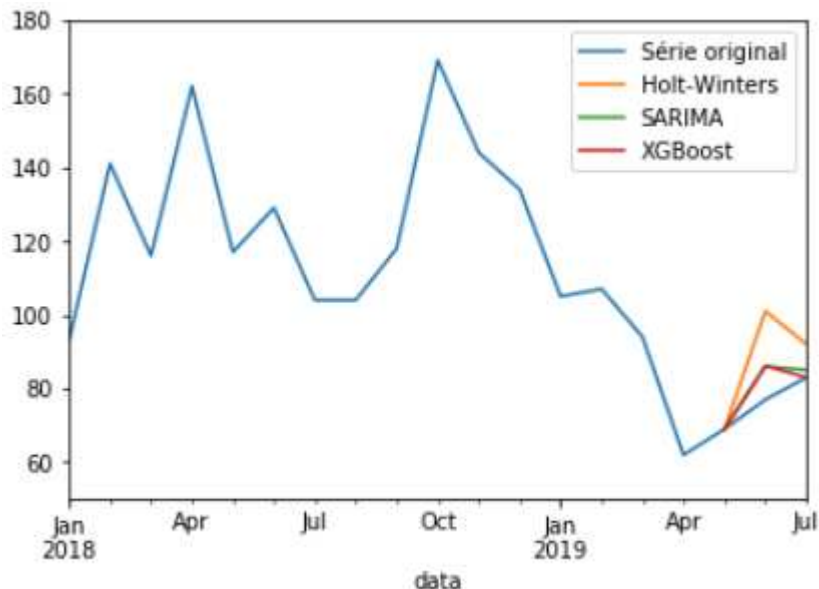


O gráfico da feature “Ruptura” mostra que há um impacto inversamente proporcional nas vendas. Os pontos que estão ao lado esquerdo do eixo indicam que valores altos de ruptura tendem a diminuir o valor de venda previsto. Da mesma forma acontece com a feature “Var Preço”, variações positivas (aumento de preço) impactam no modelo diminuindo a previsão enquanto variações negativas (queda de preço) impactam positivamente aumentando as vendas.

#### 6.4 ANÁLISE VISUAL DAS PREVISÕES OBTIDAS PARA OS ITENS

Os dados do histórico de vendas dos itens compreendem o intervalo de janeiro de 2016 até julho de 2019, sendo o volume de vendas dos últimos dois meses o que os modelos foram induzidos a prever neste trabalho. Para melhor visualização dos resultados, os gráficos com as previsões do volume de vendas e a série original de cada item foram plotados no intervalo de janeiro de 2018 a julho de 2019. A Figura 40 apresenta os resultados para o item I. Os três modelos apontaram um volume de vendas maior do que ocorreu no mês de junho, mas se aproximaram no mês de julho. Provavelmente esta superestimação na previsão deve-se que a série apresenta valores bem abaixo da média histórica justamente neste período.

Figura 40 – Previsão do volume de vendas (em unidades) para o item I.

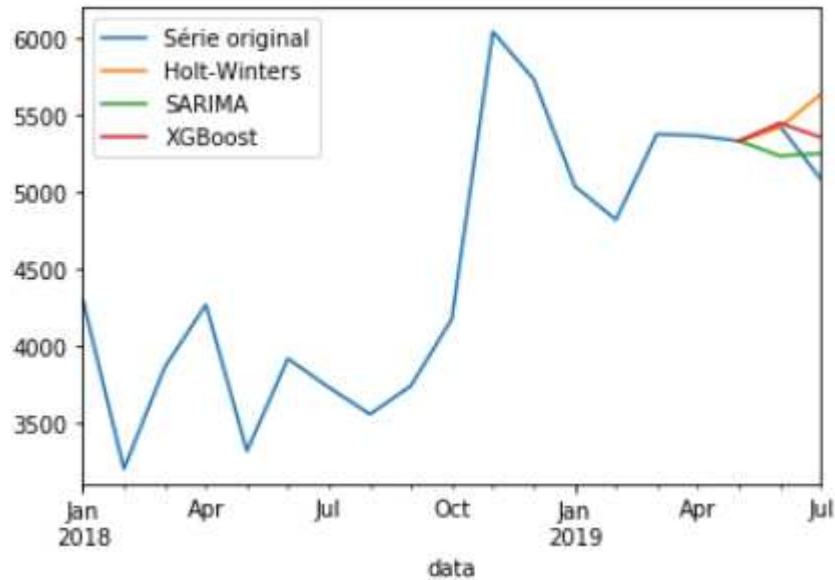


Fonte: Elaborado pelo autor (2019).

A Figura 41 apresenta os resultados para o item II. A série mostra-se bem irregular ao longo do período mostrado, porém os erros de previsão percentuais do XGBoost ficam em 2,7%

e do SARIMA 3,4%.

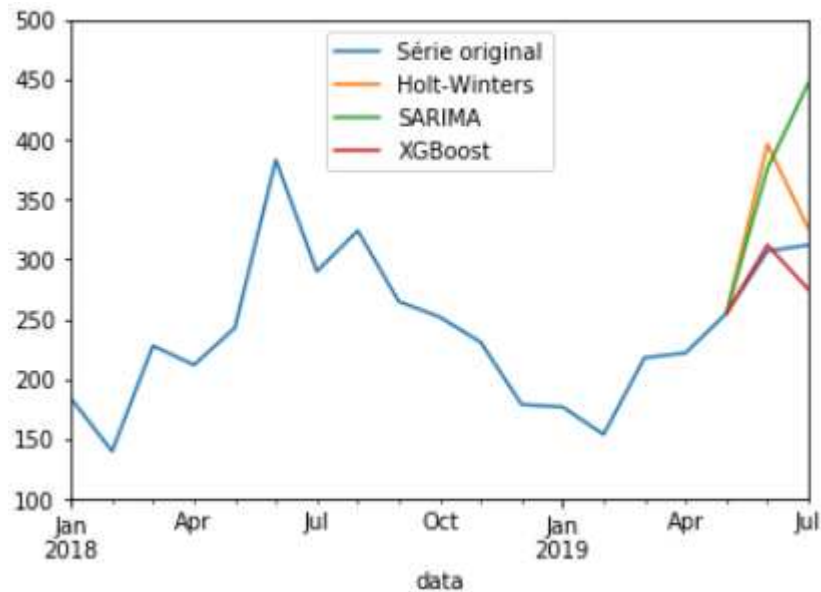
Figura 41 – Previsão do volume de vendas (em unidades) para o item II.



Fonte: Elaborado pelo autor (2019).

Os resultados para o item III são mostrados na Figura 42. Nesse caso as previsões dos modelos Holt-Winters e SARIMA ficaram muito acima do valor observado.

Figura 42 – Previsão do volume de vendas (em unidades) para o item III.

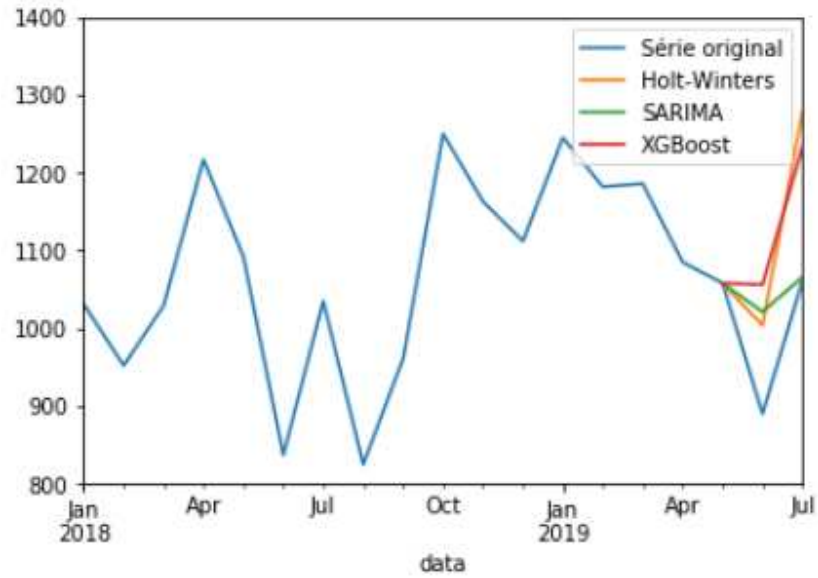


Fonte: Elaborado pelo autor (2019).

A Figura 43 retrata os resultados para o item IV. Os modelos falharam em capturar a queda no

volume de vendas que ocorreu, porém o SARIMA mostrou-se melhor que os demais modelos e obteve menor erro dentre eles.

Figura 43 – Previsão do volume de vendas (em unidades) para o item IV.



Fonte: Elaborado pelo autor (2019).

## 7 CONCLUSÕES E RECOMENDAÇÕES

O presente trabalho teve como objetivo propor um modelo de previsão da demanda de itens do setor do varejo de materiais de construção. A aplicação dos modelos apresentados seguiu os preceitos do ciclo da análise de dados, o qual serviu de linha mestra para a realização do trabalho. A definição e o refinamento da questão estão materializados nos objetivos, geral e específicos.

A etapa de coleta dos dados foi executada a partir de uma consulta em SQL diretamente do banco de dados da empresa. Dados históricos de volume de vendas, preço, estoque e características do item em um intervalo de 43 meses serviram de base para a elaboração e avaliação do modelo proposto. A análise exploratória dos dados permitiu extrair informações relevantes do conjunto de dados que contribuíram com a etapa de geração de *features*. As *features* foram geradas com base em técnicas de previsão de séries temporais, como a utilização dos *lags*. As demais *features* foram formuladas a partir da experiência do autor com a dinâmica de variação de volume de vendas do varejo observada nos dados coletados.

Técnicas de aprendizagem de máquina, utilizadas neste trabalho, necessitam de tratamento específico, como o processo de validação cruzada, que deve respeitar o caráter temporal dos dados inerente a séries temporais. O processo de tunagem dos hiperparâmetros encontrou os melhores valores dentre os especificados para a tunagem e garantiu ao modelo melhor precisão nos resultados sem que ocorresse *overfitting*.

Os resultados obtidos permitem afirmar que nenhum dos três modelos aplicados apresentou resultados superiores aos demais para os quatro itens com volume de vendas diferenciado. Porém, o modelo XGBoost gerou previsões de volume de vendas com margem de erro médio superior aos demais, tendo sido melhor, em relação ao erro, para as séries de três dos quatro itens analisados. Dos resultados conclui-se que o XGBoost fornece previsões melhores com base nas métricas avaliadas, RMSE e MAPE.

Para trabalhos futuros, recomenda-se incluir dados de todos os itens comercializados pela empresa, que perfaz um total de 45 mil SKUs. Além disso, informações quanto à exposição do item nas lojas, promoções, anúncios e outras estratégias de marketing podem tornar o modelo mais robusto visto que estas estratégias tendem a influenciar pontualmente no volume de vendas. Recomenda-se também avaliar o uso de outros métodos de previsão, em especial os de redes neurais, como o LSTM (*Long Short-Term Memory*), ou outro que venha a ganhar destaque na questão de previsão em séries temporais.

## 8 REFERÊNCIAS

BALLOU, R. H. **Gerenciamento da Cadeia de Suprimentos/Logística Empresarial**. São Paulo: Editora Bookman, 2006.

BIECEK, P.; BURZYKOWSKI, T. **Predictive Models: explore, explain and debug**. Human-Centered Interpretable Machine Learning. C. 12, 2019.

BOWERSOX, D. J.; CLOSS, D. J. **Logística Empresarial, O Processo de Integração da Cadeia de Suprimento**. São Paulo, Editora Atlas, 2009.

BARROW, D.; CRONE, S. F. **A comparison of a AdaBoost algorithms for time series forecast combination**. International Journal of Forecasting, v. 32, p. 1103-1119, 2016.

BERTO, R. M. V. S.; NAKANO, D. N. A. **Produção Científica nos Anais do Encontro Nacional de Engenharia de Produção: Um Levantamento de Métodos e Tipos de Pesquisa**. Produção, v. 9, n. 2, p. 65-76, 2000.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. International Conference on Knowledge Discovery and Data Mining, p. 785-794, 2016.

CRUZ, G. Ruptura: o consumidor não perdoa. **Especialistas em logística e supply chain**. Disponível em: <<https://www.ilos.com.br/web/ruptura-o-consumidor-nao-perdoa/>>. Acesso em: 13 de set. de 2019.

DMITRIEVSKY, M. Floresta de decisão aleatória na aprendizagem por reforço. **MQL5**. Disponível em: <<https://www.mql5.com/pt/articles/3856>>. Acesso em: 20 de jun. de 2019.

FRIES, C. E. **Underfitting vs. Overfitting** – Apêndice D. Aula ministrada na disciplina Tópicos Especiais em Pesquisa Operacional do curso de Graduação em Engenharia de Produção na Universidade Federal de Santa Catarina. Florianópolis, 2018.

GAMA, J. A. **Functional trees**. *Machine Learning*, v. 55, p. 219–250, 2004.

GRUEN, T. W.; CORSTEN, D.; BHARADWAJ, S. **Retail Out-of-Stocks: A Worldwide Examination of Causes, Rates, and Consumer Responses. Grocery Manufacturers of America.** Washington DC, 2002.

HANKE, John E.; WICHERN, Dean W. **Business Forecasting.** New Jersey: Pearson Education, Inc., 2007.

JAIN, A.; MENON, M. N.; CHANDRA, S. **Sales forecasting for retail chains.** Semantic Scholar, 2015.

KUMAR, A.; SHANKAR, R.; ALJOHANI, N. R. **A big data driven framework for demand-driven forecasting with effects of marketing-mix variables.** Industrial Marketing Management, 2019.

LACERDA, R. T. O.; ENSSLIN, L.; ENSSLIN, S. R. **Uma análise bibliométrica da literatura sobre estratégia e avaliação de desempenho.** Gestão & Produção, p. 59-78, 2012.

LEEK, J. **The Elements of Data Analytic Style: A guide for people who want to analyze data.** Leanpub, p. 98, 2015.

LUNDBERG, S.; LEE, S-I. A unified approach to interpreting model predictions. **Corneel University.** Disponível em: <<https://arxiv.org/abs/1705.07874>>. Acesso em 25 de ago. de 2019.

MA, S.; FILDES, R.; HUANG, T. **Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information.** European Journal of Operational Research, v. 249, p. 245-257, 2016.

PEDREGOSA, *et al.* Machine Learning in Python. **Scikit-learn**, p. 2825-2830, 2011.

PENG, R. D.; MATSUI, E. **The Art of Data Science: A Guide for Anyone Who Works with Data.** Lulu.com, p. 155, 2016.

ROBINZONOV, N.; TUTZ, G.; HOTHORN, T. **Boosting Techniques for Nonlinear Time Series Models**. Institut Für Statistik, 2010.

ROGOJAN, B. Ensemble Methods to Optimize Machine Learning. **Packt**. Disponível em: <<https://hub.packtpub.com/ensemble-methods-optimize-machine-learning-models/>> Acesso em: 02 de mar. De 2019.

SEAMAN, B. **Considerations of a retail forecasting practitioner**. International Journal of Forecasting, v. 34, p. 822-829, 2018.

VEEN, F. V. The neural Network Zoo. **The Asimov Institute**. Disponível em: <<https://www.asimovinstitute.org/neural-network-zoo/>>. Acesso em: 28 de ago. de 2019.

WAYNE, W. Confounding and Effect Measure Modification. **Boston University School of Public Health**. Disponível em: <[http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713\\_Confounding-EM/BS704-EP713\\_Confounding-EM2.html](http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_Confounding-EM/BS704-EP713_Confounding-EM2.html)>. Acesso em: 25 de ago. de 2019.

APÊNDICE A – BASE DE DADOS REFERENTE AO ITEM I

CODITEM	MES	Venda	Var Preço	Ruptura	Venda_lag_1	Venda_lag_2	Venda_lag_3	Venda_lag_4	Venda_lag_5	Venda_lag_6	Venda_lag_7	Venda_lag_8	Venda_lag_9	Venda_lag_10	Venda_lag_11	Venda_lag_12	Venda_lag_13	MédiaFamilia_1	Var LY
44561	1	302	0,00%	0															
44561	2	206	-7,85%	0	302														148,01
44561	3	279	5,91%	0	206	302													140,91
44561	4	235	19,44%	0	279	206	302												155,09
44561	5	214	9,68%	0	235	279	206	302											152,11
44561	6	301	-8,56%	0,012	214	235	279	206	302										159,72
44561	7	230	0,77%	0	301	214	235	279	206	302									154,19
44561	8	257	-0,78%	0,135	230	301	214	235	279	206	302								177,37
44561	9	174	-1,18%	0,15	257	230	301	214	235	279	206	302							168,60
44561	10	190	-1,20%	0,027	174	257	230	301	214	235	279	206	302						163,23
44561	11	298	0,79%	0,082	190	174	257	230	301	214	235	279	206	302					184,03
44561	12	222	-1,61%	0,226	298	190	174	257	230	301	214	235	279	206	302				187,31
44561	13	181	-7,33%	0,123	222	298	190	174	257	230	301	214	235	279	206	302			172,70
44561	14	177	0,00%	0,004	181	222	298	190	174	257	230	301	214	235	279	206	302		137,35
44561	15	191	-6,91%	0,076	177	181	222	298	190	174	257	230	301	214	235	279	206	302	123,46
44561	16	230	3,98%	0,038	191	177	181	222	298	190	174	257	230	301	214	235	279	206	127,89
44561	17	240	1,31%	0,145	230	191	177	181	222	298	190	174	257	230	301	214	235	279	148,25
44561	18	156	3,38%	0,049	240	230	191	177	181	222	298	190	174	257	230	301	214	235	160,88
44561	19	223	-6,28%	0,319	156	240	230	191	177	181	222	298	190	174	257	230	301	214	209,45
44561	20	141	2,62%	0,297	223	156	240	230	191	177	181	222	298	190	174	257	230	301	166,36
44561	21	131	-0,88%	0,249	141	223	156	240	230	191	177	181	222	298	190	174	257	230	337,57
44561	22	173	0,44%	0,179	131	141	223	156	240	230	191	177	181	222	298	190	174	257	119,20
44561	23	197	1,72%	0,114	173	131	141	223	156	240	230	191	177	181	222	298	190	174	177,42
44561	24	143	3,73%	0,008	197	173	131	141	223	156	240	230	191	177	181	222	298	190	249,18
44561	25	93	15,14%	0,209	143	197	173	131	141	223	156	240	230	191	177	181	222	298	95,46
44561	26	141	4,38%	0,117	93	143	197	173	131	141	223	156	240	230	191	177	181	222	143,05
44561	27	116	-2,77%	0,106	141	93	143	197	173	131	141	223	156	240	230	191	177	181	187,33
44561	28	162	1,03%	0,012	116	141	93	143	197	173	131	141	223	156	240	230	191	177	271,34
44561	29	117	3,63%	0	162	116	141	93	143	197	173	131	141	223	156	240	230	191	190,80
44561	30	129	-1,34%	0,03	117	162	116	141	93	143	197	173	131	141	223	156	240	230	146,76
44561	31	104	0,00%	0	129	117	162	116	141	93	143	197	173	131	141	223	156	240	169,91
44561	32	104	3,55%	0,002	104	129	117	162	116	141	93	143	197	173	131	141	223	156	116,59
44561	33	118	-0,32%	0,009	104	104	129	117	162	116	141	93	143	197	173	131	141	223	148,86
44561	34	169	-10,36%	0,033	118	104	104	129	117	162	116	141	93	143	197	173	131	141	90,94
44561	35	144	-2,19%	0,152	169	118	104	104	129	117	162	116	141	93	143	197	173	131	136,03
44561	36	134	-5,38%	0,041	144	169	118	104	104	129	117	162	116	141	93	143	197	173	152,15
44561	37	105	5,11%	0,006	134	144	169	118	104	104	129	117	162	116	141	93	143	197	162,82
44561	38	107	3,86%	0,006	105	134	144	169	118	104	104	129	117	162	116	141	93	143	139,69
44561	39	94	-2,52%	0,073	107	105	134	144	169	118	104	104	129	117	162	116	141	93	176,92
44561	40	62	-0,36%	0,025	94	107	105	134	144	169	118	104	104	129	117	162	116	141	169,04
44561	41	69	-0,73%	0,011	62	94	107	105	134	144	169	118	104	104	129	117	162	116	163,20
44561	42	77	3,85%	0,152	69	62	94	107	105	134	144	169	118	104	104	129	117	162	65,76
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	96,62
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	158,83
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	160,74
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	155,83
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	175,40
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	192,45
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	195,74
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	104,53
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	186,01
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	87,15
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	156,13
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	159,19
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	88,03
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	148,77
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	162,57
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	131,28
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	159,09
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	44,78
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	76,08
44561	43	83	8,04%	0,024	77	69	62	94	107	105	134	144	169	118	104	104	129	117	62,08



APÊNDICE B – BASE DE DADOS REFERENTE AO ITEM II

CODITEM	MES	Venda	Var Preço	Ruptura	Venda_lag_1	Venda_lag_2	Venda_lag_3	Venda_lag_4	Venda_lag_5	Venda_lag_6	Venda_lag_7	Venda_lag_8	Venda_lag_9	Venda_lag_10	Venda_lag_11	Venda_lag_12	Venda_lag_13	MédiaFamilia_1	Var LY
86882	1	1257	0,00%	0,001															
86882	2	1201	2,70%	0,025	1257													148,01	
86882	3	1157	0,00%	0,02	1201	1257												140,91	
86882	4	1223	-12,12%	0,011	1157	1201	1257											155,09	
86882	5	1721	-4,76%	0,03	1223	1157	1201	1257										152,11	
86882	6	1106	0,00%	0,109	1721	1223	1157	1201	1257									159,72	
86882	7	1495	19,23%	0	1106	1721	1223	1157	1201	1257								154,19	
86882	8	1432	-1,30%	0,002	1495	1106	1721	1223	1157	1201	1257							177,37	
86882	9	1209	1,28%	0,045	1432	1495	1106	1721	1223	1157	1201	1257						168,60	
86882	10	1697	-2,63%	0,036	1209	1432	1495	1106	1721	1223	1157	1201	1257					163,23	
86882	11	1593	1,30%	0,071	1697	1209	1432	1495	1106	1721	1223	1157	1201	1257				184,03	
86882	12	1825	1,28%	0,237	1593	1697	1209	1432	1495	1106	1721	1223	1157	1201	1257			187,31	
86882	13	1102	-1,30%	0,145	1825	1593	1697	1209	1432	1495	1106	1721	1223	1157	1201	1257		172,70	
86882	14	1332	2,53%	0,021	1102	1825	1593	1697	1209	1432	1495	1106	1721	1223	1157	1201	1257	137,35	1052,91
86882	15	2053	-14,49%	0,007	1332	1102	1825	1593	1697	1209	1432	1495	1106	1721	1223	1157	1201	127,89	1283,20
86882	16	1692	4,17%	0,02	2053	1332	1102	1825	1593	1697	1209	1432	1495	1106	1721	1223	1157	148,25	2170,11
86882	17	2246	-1,41%	0	1692	2053	1332	1102	1825	1593	1697	1209	1432	1495	1106	1721	1223	154,54	2380,97
86882	18	2329	-4,41%	0,027	2246	1692	2053	1332	1102	1825	1593	1697	1209	1432	1495	1106	1721	166,36	1443,39
86882	19	2802	12,82%	0,089	2329	2246	1692	2053	1332	1102	1825	1593	1697	1209	1432	1495	1106	161,66	3148,15
86882	20	2861	0,00%	0,084	2802	2329	2246	1692	2053	1332	1102	1825	1593	1697	1209	1432	1495	177,42	2683,92
86882	21	4581	-20,00%	0,043	2861	2802	2329	2246	1692	2053	1332	1102	1825	1593	1697	1209	1432	163,53	2415,47
86882	22	3909	-4,84%	0,016	4581	2861	2802	2329	2246	1692	2053	1332	1102	1825	1593	1697	1209	190,67	6430,07
86882	23	3026	-5,08%	0,064	3909	4581	2861	2802	2329	2246	1692	2053	1332	1102	1825	1593	1697	187,33	3669,44
86882	24	3258	9,23%	0,061	3026	3909	4581	2861	2802	2329	2246	1692	2053	1332	1102	1825	1593	190,80	3466,70
86882	25	4295	0,00%	0,012	3258	3026	3909	4581	2861	2802	2329	2246	1692	2053	1332	1102	1825	169,91	1967,30
86882	26	3202	5,80%	0,088	4295	3258	3026	3909	4581	2861	2802	2329	2246	1692	2053	1332	1102	148,86	5191,42
86882	27	3863	0,00%	0,007	3202	4295	3258	3026	3909	4581	2861	2802	2329	2246	1692	2053	1332	136,03	4935,21
86882	28	4266	0,00%	0	3863	3202	4295	3258	3026	3909	4581	2861	2802	2329	2246	1692	2053	162,82	3183,73
86882	29	3317	0,00%	0,044	4266	3863	3202	4295	3258	3026	3909	4581	2861	2802	2329	2246	1692	176,92	5662,79
86882	30	3916	2,82%	0,006	3317	4266	3863	3202	4295	3258	3026	3909	4581	2861	2802	2329	2246	153,95	3439,58
86882	31	3730	0,00%	0,003	3916	3317	4266	3863	3202	4295	3258	3026	3909	4581	2861	2802	2329	155,89	4711,31
86882	32	3556	2,74%	0,001	3730	3916	3317	4266	3863	3202	4295	3258	3026	3909	4581	2861	2802	163,20	3808,54
86882	33	3740	0,00%	0,013	3556	3730	3916	3317	4266	3863	3202	4295	3258	3026	3909	4581	2861	159,83	5693,83
86882	34	4175	0,00%	0,018	3740	3556	3730	3916	3317	4266	3863	3202	4295	3258	3026	3909	4581	160,74	3191,37
86882	35	6036	0,00%	0,009	4175	3740	3556	3730	3916	3317	4266	3863	3202	4295	3258	3026	3909	175,40	3231,91
86882	36	5727	7,59%	0,021	6036	4175	3740	3556	3730	3916	3317	4266	3863	3202	4295	3258	3026	195,74	6498,77
86882	37	5034	0,00%	0,026	5727	6036	4175	3740	3556	3730	3916	3317	4266	3863	3202	4295	3258	186,01	7549,87
86882	38	4817	0,00%	0,022	5034	5727	6036	4175	3740	3556	3730	3916	3317	4266	3863	3202	4295	156,13	3752,94
86882	39	5372	0,00%	0,012	4817	5034	5727	6036	4175	3740	3556	3730	3916	3317	4266	3863	3202	148,77	5811,39
86882	40	5362	-1,28%	0,002	5372	4817	5034	5727	6036	4175	3740	3556	3730	3916	3317	4266	3863	162,57	5932,42
86882	41	5328	4,88%	0,05	5362	5372	4817	5034	5727	6036	4175	3740	3556	3730	3916	3317	4266	159,09	4169,19
86882	42	5425	7,87%	0,016	5328	5362	5372	4817	5034	5727	6036	4175	3740	3556	3730	3916	3317	157,25	6290,16
86882	43	5080	0,00%	0,046	5425	5328	5362	5372	4817	5034	5727	6036	4175	3740	3556	3730	3916	150,09	5167,33

APÊNDICE C – BASE DE DADOS REFERENTE AO ITEM III

CODITEM	MES	Venda	Var Preço	Ruptura	Venda_lag_1	Venda_lag_2	Venda_lag_3	Venda_lag_4	Venda_lag_5	Venda_lag_6	Venda_lag_7	Venda_lag_8	Venda_lag_9	Venda_lag_10	Venda_lag_11	Venda_lag_12	Venda_lag_13	MédiaFamília_1	Var LY
74315	1	216	0,00%	0															
74315	2	139	1,03%	0	216														23,71
74315	3	154	-7,86%	0	139	216													18,89
74315	4	221	-5,56%	0	154	139	216												19,94
74315	5	338	-6,32%	0	221	154	139	216											26,60
74315	6	399	0,33%	0,038	338	221	154	139	216										36,24
74315	7	238	7,16%	0,168	399	338	221	154	139	216									44,55
74315	8	253	0,08%	0,01	238	399	338	221	154	139	216								32,89
74315	9	253	-0,62%	0,002	253	238	399	338	221	154	139	216							26,08
74315	10	197	5,62%	0,011	253	253	238	399	338	221	154	139	216						24,23
74315	11	193	5,91%	0	197	253	253	238	399	338	221	154	139	216					21,87
74315	12	186	0,61%	0,025	193	197	253	238	399	338	221	154	139	216					21,74
74315	13	164	-0,41%	0	186	193	197	253	238	399	338	221	154	139	216				17,40
74315	14	128	-0,14%	0	164	186	193	197	253	238	399	338	221	154	139	216			16,76
74315	15	137	0,14%	0,03	128	164	186	193	197	253	238	399	338	221	154	139	216		105,54
74315	16	128	-0,07%	0,196	137	128	164	186	193	197	253	238	399	338	221	154	139		141,81
74315	17	177	-1,39%	0,234	128	137	128	164	186	193	197	253	238	399	338	221	154		196,60
74315	18	342	-1,13%	0,019	177	128	137	128	164	186	193	197	253	238	399	338	221		195,76
74315	19	459	-1,94%	0,153	342	177	128	137	128	164	186	193	197	253	238	399	338		20,97
74315	20	315	-1,46%	0,01	459	342	177	128	137	128	164	186	193	197	253	238	399		208,94
74315	21	300	-4,41%	0	315	459	342	177	128	137	128	164	186	193	197	253	238		33,58
74315	22	269	10,11%	0,016	300	315	459	342	177	128	137	128	164	186	193	197	253		204,00
74315	23	253	7,46%	0,003	269	300	315	459	342	177	128	137	128	164	186	193	197		32,87
74315	24	187	-1,15%	0	253	269	300	315	459	342	177	128	137	128	164	186	193		487,93
74315	25	184	-2,09%	0	187	253	269	300	315	459	342	177	128	137	128	164	186		26,34
74315	26	140	-0,79%	0	184	187	253	269	300	315	459	342	177	128	137	128	164		315,00
74315	27	228	-2,08%	0	140	184	187	253	269	300	315	459	342	177	128	137	128		24,10
74315	28	212	-1,99%	0,007	228	140	184	187	253	269	300	315	459	342	177	128	137		233,60
74315	29	243	8,69%	0,003	212	228	140	184	187	253	269	300	315	459	342	177	128		22,63
74315	30	383	-0,95%	0,002	243	212	228	140	184	187	253	269	300	315	459	342	177		263,54
74315	31	290	-13,22%	0	383	243	212	228	140	184	187	253	269	300	315	459	342		243,82
74315	32	324	6,80%	0,021	290	383	243	212	228	140	184	187	253	269	300	315	459		19,05
74315	33	265	-0,13%	0	324	290	383	243	212	228	140	184	187	253	269	300	315		164,88
74315	34	252	3,60%	0,01	265	324	290	383	243	212	228	140	184	187	253	269	300		128,29
74315	35	231	-0,06%	0,009	252	265	324	290	383	243	212	228	140	184	187	253	269		19,05
74315	36	179	0,26%	0	231	252	265	324	290	383	243	212	228	140	184	187	253		237,01
74315	37	177	-1,76%	0,005	179	231	252	265	324	290	383	243	212	228	140	184	187		23,05
74315	38	154	-1,32%	0,007	177	179	231	252	265	324	290	383	243	212	228	140	184		21,89
74315	39	218	0,40%	0,001	154	177	179	231	252	265	324	290	383	243	212	228	140		18,71
74315	40	222	0,20%	0,014	218	154	177	179	231	252	265	324	290	383	243	212	228		170,74
74315	41	255	-1,40%	0,022	222	218	154	177	179	231	252	265	324	290	383	243	212		176,13
74315	42	307	0,00%	0,032	255	222	218	154	177	179	231	252	265	324	290	383	243		19,48
74315	43	312	6,20%	0,064	307	255	222	218	154	177	179	231	252	265	324	290	383		134,67
74315																			17,32
74315																			250,80
74315																			202,70
74315																			21,55
74315																			202,70
74315																			254,46
74315																			22,27
74315																			401,91
74315																			24,73
74315																			401,91
74315																			232,45

APÊNDICE D – BASE DE DADOS REFERENTE AO ITEM IV

CODITEM	MES	Venda	Var Preço	Ruptura	Venda_lag_1	Venda_lag_2	Venda_lag_3	Venda_lag_4	Venda_lag_5	Venda_lag_6	Venda_lag_7	Venda_lag_8	Venda_lag_9	Venda_lag_10	Venda_lag_11	Venda_lag_12	Venda_lag_13	MédiaFamilia_1	Var LY
129794	1	694	0,00%	0,176															
129794	2	617	3,21%	0,124	694														69,47
129794	3	841	-5,16%	0,1	617	694													65,91
129794	4	1041	-0,40%	0,024	841	617	694												74,08
129794	5	725	24,17%	0,041	1041	841	617	694											70,86
129794	6	759	-11,82%	0,033	725	1041	841	617	694										67,44
129794	7	993	-6,67%	0,007	759	725	1041	841	617	694									67,01
129794	8	886	-0,91%	0,02	993	759	725	1041	841	617	694								80,51
129794	9	701	-0,18%	0,055	886	993	759	725	1041	841	617	694							76,29
129794	10	728	-3,00%	0,191	701	886	993	759	725	1041	841	617	694						71,16
129794	11	980	-0,76%	0,06	728	701	886	993	759	725	1041	841	617	694					78,74
129794	12	584	2,58%	0,221	980	728	701	886	993	759	725	1041	841	617	694				82,42
129794	13	812	-7,74%	0,171	584	980	728	701	886	993	759	725	1041	841	617	694			72,31
129794	14	911	-1,00%	0,064	812	584	980	728	701	886	993	759	725	1041	841	617	694		68,22
129794	15	865	14,55%	0,106	911	812	584	980	728	701	886	993	759	725	1041	841	617	694	721,91
129794	16	1012	-7,35%	0,116	865	911	812	584	980	728	701	886	993	759	725	1041	841	617	68,34
129794	17	917	7,48%	0,154	1012	865	911	812	584	980	728	701	886	993	759	725	1041	841	77,79
129794	18	932	-6,72%	0,028	917	1012	865	911	812	584	980	728	701	886	993	759	725	1041	70,80
129794	19	1223	0,72%	0,03	932	917	1012	865	911	812	584	980	728	701	886	993	759	725	73,41
129794	20	961	-3,16%	0,08	1223	932	917	1012	865	911	812	584	980	728	701	886	993	759	78,67
129794	21	779	4,78%	0,193	961	1223	932	917	1012	865	911	812	584	980	728	701	886	993	1091,22
129794	22	1367	-2,91%	0,112	779	961	1223	932	917	1012	865	911	812	584	980	728	701	886	66,52
129794	23	1022	8,65%	0,105	1367	779	961	1223	932	917	1012	865	911	812	584	980	728	701	76,32
129794	24	841	9,62%	0,076	1022	1367	779	961	1223	932	917	1012	865	911	812	584	980	728	809,00
129794	25	1031	-13,48%	0,065	841	1022	1367	779	961	1223	932	917	1012	865	911	812	584	980	74,81
129794	26	952	-7,72%	0,093	1031	841	1022	1367	779	961	1223	932	917	1012	865	911	812	584	77,86
129794	27	1029	8,11%	0,119	952	1031	841	1022	1367	779	961	1223	932	917	1012	865	911	812	609,03
129794	28	1217	0,00%	0,002	1029	952	1031	841	1022	1367	779	961	1223	932	917	1012	865	911	75,07
129794	29	1092	-5,71%	0,026	1217	1029	952	1031	841	1022	1367	779	961	1223	932	917	1012	865	69,62
129794	30	837	12,91%	0,004	1092	1217	1029	952	1031	841	1022	1367	779	961	1223	932	917	1012	903,93
129794	31	1035	-1,74%	0,014	837	1092	1217	1029	952	1031	841	1022	1367	779	961	1223	932	917	78,23
129794	32	825	2,62%	0,024	1035	837	1092	1217	1029	952	1031	841	1022	1367	779	961	1223	932	61,94
129794	33	961	3,13%	0,024	825	1035	837	1092	1217	1029	952	1031	841	1022	1367	779	961	1223	1109,86
129794	34	1250	-15,72%	0,024	961	825	1035	837	1092	1217	1029	952	1031	841	1022	1367	779	961	66,10
129794	35	1163	14,22%	0,014	1250	961	825	1035	837	1092	1217	1029	952	1031	841	1022	1367	779	74,01
129794	36	1112	-5,97%	0,021	1163	1250	961	825	1035	837	1092	1217	1029	952	1031	841	1022	1367	813,27
129794	37	1245	-2,41%	0,018	1112	1163	1250	961	825	1035	837	1092	1217	1029	952	1031	841	1022	66,76
129794	38	1182	-1,30%	0	1245	1112	1163	1250	961	825	1035	837	1092	1217	1029	952	1031	841	70,31
129794	39	1186	0,97%	0,008	1182	1245	1112	1163	1250	961	825	1035	837	1092	1217	1029	952	1031	1686,38
129794	40	1085	5,20%	0,011	1186	1182	1245	1112	1163	1250	961	825	1035	837	1092	1217	1029	952	71,59
129794	41	1058	0,30%	0,02	1085	1186	1182	1245	1112	1163	1250	961	825	1035	837	1092	1217	1029	934,53
129794	42	890	10,63%	0,005	1058	1085	1186	1182	1245	1112	1163	1250	961	825	1035	837	1092	1217	73,68
129794	43	1061	-7,31%	0,043	890	1058	1085	1186	1182	1245	1112	1163	1250	961	825	1035	837	1092	957,03
																			80,44
																			1363,22
																			70,08
																			1149,60
																			59,27
																			1277,60
																			70,10
																			1402,68
																			58,08
																			973,56
																			61,94
																			810,94
																			64,88
																			1100,54