UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Ferramental matemático e computacional para apoio a gestão de pequenos supermercados

Cristian Júlio de Barros

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Cristian Júlio de Barros

Ferramental matemático e computacional para apoio a gestão de pequenos supermercados

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestre – Mestrado Profissional em Matemática,
Estatística e Computação Aplicadas à Indústria.
VERSÃO REVISADA

Área de Concentração: Matemática, Estatística e
Computação

Orientador: Prof. Dr. Luis Gustavo Nonato

USP – São Carlos
Janeiro de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

B277f Barros, Cristian Julio de
 Ferramental matemático e computacional para
apoio a gestão de pequenos supermercados / Cristian
Julio de Barros; orientador Luis Gustavo Nonato. --
São Carlos, 2020.
 126 p.

 Dissertação (Mestrado - Programa de Pós-Graduação
em Mestrado Profissional em Matemática, Estatística
e Computação Aplicadas à Indústria) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2020.

 1. Matemática. 2. Estatística. 3. Computação. 4.
Varejo. 5. Aprendizado de Máquina. I. Nonato, Luis
Gustavo, orient. II. Título.

Cristian Júlio de Barros

**Mathematical and computational tools to support the
management of small supermarkets**

Master dissertation submitted to the Institute of
Mathematics and Computer Sciences – ICMC-USP, in
partial fulfillment of the requirements for the degree of
the Master – Professional Masters in Mathematics,
Statistics and Computing Applied to Industry. *FINAL
VERSION*

Concentration Area: Mathematics, Statistics and
Computing

Advisor: Prof. Dr. Luis Gustavo Nonato

USP – São Carlos
January 2020

*Este trabalho é dedicado a todos os empreendedores,
trabalhadores e pesquisadores que com garra e determinação
contribuem para o progresso do país.*

AGRADECIMENTOS

Sou grato à minha família e amigos pelo apoio recebido quando iniciei este mestrado e incentivaram a superar as dificuldades encontradas nesta longa jornada.

As empresas CTIS Tecnologia e CAIXA ECONÔMICA FEDERAL que por meio de seus gestores autorizaram a minha participação neste mestrado.

Ao Orientador Professor Dr. Luis Gustavo Nonato, e a todos os demais professores do MECAL, pela paciência e apoio na execução de todas as atividades.

*“Se a aparência e a essência das coisas coincidissem,
a ciência seria desnecessária.”
(Karl Marx)*

RESUMO

CRISTIAN, C. B. **Ferramental matemático e computacional para apoio a gestão de pequenos supermercados**. 2020. 126 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

No Brasil em 2015 existiam 1.193.090 empresas varejistas de pequeno porte (fonte: IBGE – Pesquisa anual de Comércio), os supermercados são empresas de varejo e quando são de pequeno porte na maioria das vezes padecem de uma gestão empírica, familiar e não profissional. Estes são fatores que contribuem para os baixos resultados operacionais detectados nos estabelecimentos de menor faturamento (BADIN, 1997). Nas últimas décadas ocorreram a massificação do acesso aos computadores, a disseminação da internet e por último a obrigatoriedade da emissão de notas fiscais eletrônicas, tornando obrigatória a informatização de todos os pequenos supermercados e a consequente adoção de ERP's adequados as necessidades do negócio. A existência desta infra-estrutura de TI possibilita que sejam utilizadas métodos de cunho matemático e computacional que possam ajudar o proprietário a obter melhores resultados no âmbito financeiro e/ou operacional.

Assim o objetivo deste projeto é identificar as ferramentas que possam ser executadas na infra-estrutura computacional de um pequeno supermercado e a elaboração de um roteiro de processos que sejam compreensíveis ao gestor e que se apresente como uma ferramenta de melhoria de resultados operacionais e financeiros para o pequeno varejo supermercadista.

Palavras-chave: Matemática, Estatística, Computação, Varejo, Aprendizado de Máquina.

ABSTRACT

CRISTIAN, C. B. **Mathematical and computational tools to support the management of small supermarkets**. 2020. 126 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

In Brazil in 2015 there were 1,193,090 small retail companies (source: IBGE - Annual Survey of Commerce), supermarkets are retail companies and when they are small they usually suffer from an empirical, family and non-profit management. professional. These are factors that contribute to the low quality results present in most of those establishments. (Badin page 79). In the last decades, the massification of access to computers, the dissemination of the Internet and, lastly, the compulsory issuance of electronic invoices, have made mandatory the digitization of all small supermarkets and consequent adoption of ERPs adequate to the needs of the business. The existence of this IT infrastructure makes it possible to use mathematical and computational methods that can help the owner achieve better financial and / or operational results.

Thus the objective of this project is to identify the tools that can be employed in the computational infrastructure of a small supermarket and the elaboration of a script of processes that are understandable to the manager and that presents itself as a tool of improvement of operational and financial results to the small retail supermarket.

Keywords: Mathematics, Statistic, Computing, Retail, Machine Learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Participação das categorias mais vendidas nos Supermercados	30
Figura 2 – Fluxograma do processo de engenharia de dados	45
Figura 3 – Produtos cadastrados erroneamente	50
Figura 4 – Reclassificação dos produtos	50
Figura 5 – Visão geral dos dados analisados	57
Figura 6 – Estrutura dos dados	58
Figura 7 – Faturamento mensal deflacionado	59
Figura 8 – Indicadores de vendas - participação mensal	59
Figura 9 – Indicadores de vendas - variação percentual	60
Figura 10 – Faturamento anual: Realizado X Ajustado pela inflação	61
Figura 11 – (%) do faturamento por hora do dia	61
Figura 12 – Participação dos indicadores de venda por dia da semana	62
Figura 13 – Participação dos indicadores de venda por dia do mês	63
Figura 14 – Participação dos indicadores de venda por mês do ano	64
Figura 15 – Participação das vendas mensais por departamento	65
Figura 16 – Participação da quantidade de unidades vendidas por departamento	65
Figura 17 – Participação da quantidade de vendas por departamento	66
Figura 18 – Visão geral - Indicadores de vendas por departamento	67
Figura 19 – Variação dos indicadores de vendas por departamento	67
Figura 20 – Vendas mensais por departamento e total	68
Figura 21 – Participação das vendas por seção nas vendas totais do período	69
Figura 22 – Participação das quantidades de unidades vendidas por seção no total de unidades vendidas no período	69
Figura 23 – Participação das quantidades de vendas por seção no total de vendas do período	70
Figura 24 – Treemap dos indicadores de vendas	71
Figura 25 – TOP 3 - seção mais relevantes	72
Figura 26 – BOTTOM 3 - Seções menos relevantes	73
Figura 27 – Resumo dos dados por grupo de produtos	74
Figura 28 – Visão geral dos grupos	75
Figura 29 – Variação das vendas por GRUPO	76
Figura 30 – Treemap dos indicadores de vendas por grupo	77
Figura 31 – TOP 3 - Grupos mais relevantes	78
Figura 32 – BOTTOM 3 - Grupos menos relevantes	79

Figura 33 – Quantidade de produtos comercializados no mês	80
Figura 34 – Quantidade de produtos não comercializados no mês	81
Figura 35 – Lista dos 30 principais produtos	81
Figura 36 – Lista dos 30 produtos menos relevantes	82
Figura 37 – RMSE da Previsão de vendas mensal: Melhores algoritmos	91
Figura 38 – RMSE da Previsão de vendas mensal: Melhores Modelos Stacking	91
Figura 39 – RMSE da Previsão de vendas mensal: Melhores Modelos Blending	92
Figura 40 – RMSE da Previsão de vendas mensal (Drugstore): Melhores Modelos Stacking	93
Figura 41 – RMSE da Previsão de vendas mensal (Não Perecíveis): Melhores Modelos Stacking	93
Figura 42 – RMSE da Previsão de vendas mensal (Perecíveis): Melhores Modelos Stacking	93
Figura 43 – (%) de desvio absoluto da Previsão de vendas mensal (Drugstore)	94
Figura 44 – (%) de desvio absoluto da Previsão de vendas mensal (Não Perecíveis)	94
Figura 45 – (%) de desvio absoluto da Previsão de vendas mensal (Perecíveis)	94
Figura 46 – RMSE da Previsão de vendas semanais: Melhores algoritmos	95
Figura 47 – RMSE da Previsão de vendas semanais: Melhores modelos Stacking	95
Figura 48 – RMSE da Previsão de vendas semanais: Melhores modelos Blending	96
Figura 49 – RMSE Previsão de vendas semanais (Drugstore): Melhores Modelos Stacking	97
Figura 50 – RMSE Previsão de vendas semanais (Departamento Não Perecíveis): Melho- res Modelos Stacking	97
Figura 51 – RMSE Previsão de vendas semanais (Perecíveis): Melhores Modelos Stacking	98
Figura 52 – (%) de desvio absoluto da Previsão de vendas semanais (Drugstore)	98
Figura 53 – (%) de desvio absoluto da Previsão de vendas semanais (Departamento Não Perecíveis)	98
Figura 54 – (%) de desvio absoluto da Previsão de vendas semanais (Perecíveis)	99
Figura 55 – RMSE da previsão de vendas diárias: Melhores algoritmos	100
Figura 56 – RMSE da previsão de vendas diárias: Melhores modelos Stacking	100
Figura 57 – RMSE da previsão de vendas diárias: Melhores modelos Blending	100
Figura 58 – RMSE da Previsão de vendas semanais (Drugstore): Melhores Modelos Stacking	102
Figura 59 – RMSE da Previsão de vendas semanais (Não Perecíveis): Melhores Modelos Stacking	102
Figura 60 – RMSE da Previsão de vendas semanais (Perecíveis): Melhores Modelos Stacking	102
Figura 61 – (%) de desvio absoluto da Previsão de vendas semanais (Drugstore): Melho- res Modelos Stacking	103
Figura 62 – (%) de desvio absoluto da Previsão de vendas semanais (Não Perecíveis): Melhores Modelos Stacking	103

Figura 63 – (%) de desvio absoluto da Previsão de vendas semanais (Perecíveis): Melhores Modelos Stacking	103
Figura 64 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:mensal) . . .	105
Figura 65 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:semanal) . .	105
Figura 66 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:diária)	105
Figura 67 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:mensal)	106
Figura 68 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:semanal)	106
Figura 69 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:diária)	106
Figura 70 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)	108
Figura 71 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal)	108
Figura 72 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)	108
Figura 73 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)	109
Figura 74 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal)	109
Figura 75 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)	109
Figura 76 – RMSE Previsão de demanda - MUSSARELA (periodicidade:mensal)	111
Figura 77 – RMSE Previsão de demanda - MUSSARELA (periodicidade:semanal) . . .	111
Figura 78 – RMSE Previsão de demanda - MUSSARELA (periodicidade:diária)	111
Figura 79 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:mensal)	112
Figura 80 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:semanal)	112
Figura 81 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:diária)	112
Figura 82 – RMSE Previsão de preço - FRANGO FILET (periodicidade:mensal)	114
Figura 83 – RMSE Previsão de preço - FRANGO FILET (periodicidade:semanal) . . .	114
Figura 84 – (%) de desvio absoluto da Previsão de preço - FRANGO FILET (periodicidade:mensal)	115
Figura 85 – (%) de desvio absoluto da Previsão de preço - FRANGO FILET (periodicidade:semanal)	115
Figura 86 – RMSE Previsão de preço - BATATA LAVADA (periodicidade:mensal) . . .	116

Figura 87 – RMSE Previsão de preço - BATATA LAVADA (periodicidade:semanal) . . .	117
Figura 88 – (%) de desvio absoluto da Previsão de preço - BATATA LAVADA (periodicidade:mensal)	117
Figura 89 – (%) de desvio absoluto da Previsão de preço - BATATA LAVADA (periodicidade:semanal)	117
Figura 90 – RMSE Previsão de preço - BANANA NANICA (periodicidade:mensal) . . .	118
Figura 91 – RMSE Previsão de preço - BANANA NANICA (periodicidade:semanal) . .	119
Figura 92 – (%) de desvio absoluto da Previsão de preço - BANANA NANICA (periodicidade:mensal)	119
Figura 93 – (%) de desvio absoluto da Previsão de preço - BANANA NANICA (periodicidade:semanal)	119

LISTA DE ABREVIATURAS E SIGLAS

ABRAS	Associação Brasileira de Supermercados
AMBIMA	Associação Brasileira das Entidades dos Mercados Financeiros e de Capitais.
AM	Aprendizado de Máquina
APAS	Associação Paulista de Supermercados
BD	Banco de Dados
DW	Data Warehouse
ERP	Enterprise Resource Planning
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMC	Instituto de Ciências Matemáticas e de Computação
IDH	Índice de Desenvolvimento Humano
INPC	Índice Nacional de Preços ao Consumidor
IPCA	Índice Nacional de Preços ao Consumidor - Amplo
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors Algorithm
MD	Mineração de Dados
MECAI	Matemática Estatística e Computação Aplicados a Indústria
MLP	Multilayer Perceptron
NN	Neural Network
OLS	Ordinal least Square
PCA	Principal Component Analysis
PDV	Ponto-de-Venda
SEBRAE	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas

SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
SVM	Suporte Vector Machine
TAE	Teoria do Aprendizado Estatístico
TI	Tecnologia da Informação
USP	Universidade de São Paulo

LISTA DE TABELAS

Tabela 1 – Divisão de comércio e faixas de pessoal ocupado	30
Tabela 2 – Evolução administrativa do SuperX	38
Tabela 3 – Lista de atributos disponibilizados pelo SuperX	48
Tabela 4 – Lista de atributos selecionados para o SuperX	49
Tabela 5 – Lista de artefatos <i>codigo fonte – EngenhariadeDados</i>	53
Tabela 6 – Níveis hierárquicos da informação	55
Tabela 7 – Atributos derivados de sumarizações e cálculos	56
Tabela 8 – Resumo dos dados por seção	68
Tabela 9 – Variação das vendas por seção	70
Tabela 10 – Lista de artefatos - Código fonte	83
Tabela 11 – Parametros da rotina de pré-tratamento dos dados.	87
Tabela 12 – Dimensão temporal	88
Tabela 13 – Parametros da rotina de avaliação de modelos.	88
Tabela 14 – Conjuntos de dados	90
Tabela 15 – RMSE - Toda a loja com periodicidade mensal.	91
Tabela 16 – RMSE - Drugstore com periodicidade mensal.	92
Tabela 17 – RMSE - Não perecíveis com periodicidade mensal.	92
Tabela 18 – RMSE - Perecíveis com periodicidade mensal.	92
Tabela 19 – RMSE - Toda a loja com periodicidade semanal.	95
Tabela 20 – RMSE - Drugstore com periodicidade semanal.	96
Tabela 21 – RMSE - Não perecíveis com periodicidade semanal.	96
Tabela 22 – RMSE - Perecíveis com periodicidade semanal.	96
Tabela 23 – RMSE - Toda a loja com periodicidade diária.	99
Tabela 24 – RMSE - Drugstore com periodicidade diária.	101
Tabela 25 – RMSE - Não perecíveis com periodicidade diária.	101
Tabela 26 – RMSE - Perecíveis com periodicidade diária.	101
Tabela 27 – RMSE - PÃO FRANCES (periodicidade:mensal)	104
Tabela 28 – RMSE - PÃO FRANCES (periodicidade:semanal)	104
Tabela 29 – RMSE - PÃO FRANCES (periodicidade:diária)	104
Tabela 30 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)	107
Tabela 31 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal) . .	107
Tabela 32 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)	107
Tabela 33 – RMSE - MUSSARELA (periodicidade:mensal)	110

Tabela 34 – RMSE - MUSSARELA (periodicidade:semanal)	110
Tabela 35 – RMSE - MUSSARELA (periodicidade:diária)	110
Tabela 36 – RMSE - Previsão de preço - FRANGO FILET (periodicidade:mensal) . . .	113
Tabela 37 – RMSE - Previsão de preço - FRANGO FILET (periodicidade:semanal) . . .	113
Tabela 38 – RMSE - Previsão de preço - BATATA LAVADA (periodicidade:mensal) . .	116
Tabela 39 – RMSE - Previsão de preço - BATATA LAVADA (periodicidade:semanal) . .	116
Tabela 40 – RMSE - Previsão de preço - BANANA NANICA (periodicidade:mensal) . .	118
Tabela 41 – RMSE - Previsão de preço - BANANA NANICA (periodicidade:semanal) .	118
Tabela 42 – Lista de artefatos <i>codigofonte – AnaliseMultivariada</i>	120

SUMÁRIO

1	INTRODUÇÃO	27
1.1	Evolução histórica do varejo	27
1.2	Definições e conceitos	28
1.3	Tipos de lojas	28
1.4	Determinantes da evolução do varejo	29
1.5	Estrutura atual do varejo no Brasil	29
1.6	Matemática utilizada pelo varejo	31
2	OBJETIVOS	35
2.1	Objetivos específicos	35
2.2	Estrutura do trabalho	36
3	METODOLOGIA DE AQUISIÇÃO DOS DADOS	37
3.1	SuperX	37
3.2	Definição de fases e processos	39
3.2.1	<i>Recepção</i>	39
3.2.2	<i>Limpeza dos dados</i>	39
3.2.3	<i>Valores ausentes</i>	39
3.2.4	<i>Valores fora de padrão</i>	40
3.2.5	<i>Dados inconsistentes</i>	40
3.2.6	<i>Integração de dados</i>	41
3.2.7	<i>Seleção dos dados</i>	41
3.2.8	<i>Transformação dos dados</i>	42
3.3	Origem dos dados	42
3.4	Processamento de dados	43
4	ENGENHARIA DE DADOS	45
4.1	Aquisição dos dados	46
4.1.1	<i>Infraestrutura</i>	46
4.1.2	<i>Hacking</i>	46
4.1.3	<i>Leitura de dados</i>	46
4.2	Seleção dos dados	47
4.3	Seleção de atributos	47
4.4	Limpeza dos dados selecionados	48

4.5	Integração dos dados	50
4.6	Transformação dos dados	51
4.7	Pré-processamento	52
4.8	Observações	53
5	ESTATÍSTICA DESCRITIVA	55
5.1	Análise exploratória de dados	57
5.2	Análise dos dados	58
5.2.1	<i>Faturamento mensal</i>	<i>58</i>
5.2.2	<i>Faturamento sazonal</i>	<i>60</i>
5.3	Análise temporal dos dados por departamento	64
5.3.1	<i>Dados gerais</i>	<i>64</i>
5.3.2	<i>Comportamento dos departamentos - relevância</i>	<i>67</i>
5.4	Análise temporal dos dados por seção	68
5.4.1	<i>Dados gerais</i>	<i>68</i>
5.4.2	<i>Principais seções</i>	<i>72</i>
5.4.3	<i>Seções sem relevância</i>	<i>73</i>
5.5	Análise temporal dos dados por grupo	74
5.5.1	<i>Dados gerais</i>	<i>74</i>
5.5.2	<i>Principais grupos</i>	<i>77</i>
5.5.3	<i>Grupos sem relevância</i>	<i>79</i>
5.6	Análise temporal dos dados por produto	80
5.6.1	<i>Visão geral dos produtos</i>	<i>80</i>
5.6.2	<i>Principais produtos comercializados</i>	<i>81</i>
5.6.3	<i>Produtos sem relevância no resultado</i>	<i>82</i>
5.7	Agrupamento de produtos	83
5.8	Observações	83
6	APRENDIZADO DE MÁQUINA	85
6.1	Conceitos básicos	85
6.2	Motor de Modelos	86
6.3	Previsão de vendas	90
6.3.1	<i>Mensal</i>	<i>90</i>
6.3.2	<i>Semanal</i>	<i>94</i>
6.3.3	<i>Diário</i>	<i>99</i>
6.4	Previsão de demanda	103
6.4.1	<i>Produto: PÃO FRANCES</i>	<i>104</i>
6.4.2	<i>Produto: REFR COCA-COLA RETORNÁVEL</i>	<i>107</i>
6.4.3	<i>Produto: MUSSARELA</i>	<i>110</i>
6.5	Previsão de preço	113

6.5.1	<i>Produto: FRANGO FILET</i>	113
6.5.2	<i>Produto: BATATA LAVADA</i>	116
6.5.3	<i>Produto: BANANA NANICA</i>	118
6.6	Observações	120
7	CONCLUSÃO E TRABALHOS FUTUROS	121
	REFERÊNCIAS	125

INTRODUÇÃO

1.1 Evolução histórica do varejo

Os registros mais antigos da existência de atacadistas e varejistas foram encontrados juntamente com as relíquias de comércio em antigas ruínas. As lojas eram numerosas em Roma e nas principais cidades do império, porém a queda do Império Romano causou a desestruturação de toda a organização de varejo na época (BADIN, 1997).

Durante a idade moderna as lojas de mercadorias gerais (*general stores*) se tornaram tradicionais nos Estados Unidos e na Inglaterra, também chamadas de empórios e armazéns no Brasil. Estas lojas comercializavam os mais variados produtos, como alimentos, tecidos, pólvora, armas, dentre outros, criando assim um importante papel social por atuarem como centro de reuniões do setor rural (BADIN, 1997).

As lojas de departamento, lojas especializadas em vendas pelo correio postal e as redes de lojas apareceram no início do século XX, fazendo emergir importantes varejistas americanos como Marshall Field, John Wanamaker e Stewart. Os *shoppings centers* começaram a se desenvolver, principalmente por oferecerem amplos estacionamentos e uma série de outras atividades (BADIN, 1997).

No caso específico do Brasil, Portugal controlava todas as ações varejistas que ocorriam em território nacional no período colonial. As companhias de comércio tinham como objetivo atingir economias de escala e tiveram importante atuação no mercado brasileiro. Enquanto a exportação era a base do comércio, internamente o país enfrentava dificuldades, atribuídas por alguns historiadores à própria centralização de decisões em Lisboa, agravadas por problemas de infra-estrutura interna, dificuldades de transporte e de comunicação, a distância entre os locais de comercialização, além das várias leis para manter o monopólio dos portugueses na comercialização (BADIN, 1997).

1.2 Definições e conceitos

Existem diferentes definições de varejo para a comercialização direta ao consumidor final. Será considerada a visão de [Levy \(2014\)](#) como mais adequada e abrangente, esta visão define a atividade de varejo como o conjunto de operações de negócios que adiciona valor a produtos e serviços vendidos para uso pessoal ou familiar dos consumidores.

A atividade varejista envolve também a venda de serviços como hospedagem em hotel, um exame médico, um corte de cabelo, locação de automóveis. Esta abordagem justifica as crescentes opções de produtos e serviços ofertados aos consumidores, de forma que estes empreguem o mínimo de recursos para satisfazer as suas necessidades e desejos ([LEVY, 2014](#)).

Os diferentes tipos de varejos se encarregam de atividades ou realizam funções que acrescentam valor aos produtos e serviços vendidos ao consumidor, e podem ser definidas como: prover sortimentos ou serviços, desmembrar grandes volume e manter estoques. Elas são utilizadas com maior ou menor ênfase em função da estratégia empregada pela organização. Lojas de descontos têm como enfoque a comercialização de produtos com baixos preços e podem eventualmente deixar de oferecer alguns serviços ao consumidor, enquanto lojas especializadas trabalham com menor variedade de produtos, mas com grande variedade de marcas ([BADIN, 1997](#)).

1.3 Tipos de lojas

Os supermercados podem ser classificados de acordo com seu porte, faturamento, abrangência regional, origem do capital, entre outros. Neste trabalho foi utilizada a seguinte classificação: ([BADIN, 1997](#))

- a) Hipermercados: São grandes áreas de auto-serviço (pegue e carregue), com extensa variedade de produtos alimentícios e bebidas. Mais de 50% do seu faturamento são obtidos com itens como roupas, acessórios de carros, artigos esportivos e residenciais. Focam na eficiência e qualidade dos produtos, com preços razoáveis e atendimento pessoal nos departamentos mais especializados. No Brasil são considerados hipermercados as lojas com mais de 5.000 m^2 ;
- b) Supermercados: São lojas de auto-serviço com alguns setores (padaria, açougue, rotisserie) oferecendo serviços de corte, pesagem, embalagem e até mesmo agendamento, além de linha completa de itens alimentícios e não-alimentícios;
- c) Lojas de Conveniência: Oferecem pequena variedade de itens para atender um público específico que procura os produtos 24 horas por dia. São relativamente pequenas, normalmente de fácil acesso e compra rápida;
- d) Lojas de Especialidades: São os açougues, as lojas de bebidas, padarias, lojas de laticínios e boutiques de alimentos. São pequenos entrepostos, com grande especialização

e prestação de serviços.

1.4 Determinantes da evolução do varejo

Existem alguns fatores que geraram e estão gerando turbulência na organização do varejo. Fatores demográficos, necessidades variadas, consumidor com comportamento diversificado e em constante mudança, somados às novidades advindas do mundo digital e oportunidades na localização dos pontos de venda, exigem uma adaptação para criar vantagens competitivas. Segundo [Levy \(2014\)](#), as organizações de varejo devem criar planos estratégicos relacionados com:

- a) A natureza dinâmica dos competidores e consumidores em mercados de varejo;
- b) Como os varejistas desenvolvem estratégias para construir vantagens competitivas e identificar oportunidades de crescimento;
- c) Abordagens de análise de performance financeira de longo prazo;
- d) O papel da localização, estrutura organizacional e gerenciamento de recursos humanos, distribuição e sistemas de informação, relacionamento com vendedores, e serviços ao cliente como fatores de desenvolvimento de vantagens estratégicas.

1.5 Estrutura atual do varejo no Brasil

As primeiras unidades de supermercado no Brasil datam de 1954. Na década de 70 foram inauguradas as primeiras lojas de hipermercado, sendo o Carrefour inaugurado em 1974 o primeiro do gênero no Brasil ([ABRAS, 1993](#)).

Foi neste período que se iniciou o processo de consolidação das redes de supermercados e a profissionalização da gestão provocando um distanciamento das práticas administrativas adotadas pelas grandes redes em comparação aos pequenos supermercados isolados ([FILHO, 2003](#)).

A Fundação Getulio Vargas (FGV) publicou um amplo estudo elaborado pelo IBGE, com base em uma grande quantidade de informações sobre o setor de varejo, este dados foram apresentados pela ABRAS e seu congresso anual realizado em 2012, com destaque para os dados sobre a participação das categorias no faturamento do supermercado exibidos na Figura 1. ([GVCEV, 2012](#))

A Associação Brasileira de Supermercados (ABRAS), e a Associação Paulista de Supermercados (APAS) entidades que representam o setor em nível nacional e estadual respectivamente, mantêm vários estudos e indicadores, porém não possuem um foco voltado ao pequeno supermercadista, e o Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE), res-

Figura 1 – Participação das categorias mais vendidas nos Supermercados

Fonte: (ABRAS) (2011)

ponsável por políticas de apoio e desenvolvimento do pequeno negócio, é generalista e também não possui foco específico neste setor.

O IBGE realizou estudo denominado “Demografia das Empresas – 2015” realizado a partir dos dados do censo econômico de 2015, destaca a importância do varejo para a economia do Brasil. Os dados estão disponíveis para consulta pública na internet (IBGE, 2015), com interface amigável e fácil extração e visualização dos dados por parte dos usuários. A pesquisa pode ser realizada selecionando as variáveis de interesse, os níveis de agrupamento e as respectivas funções de agregação (média, soma, mínimo, máximo, desvio-padrão), como pode ser visto na tabela 1:

Tabela 1 – Divisão de comércio e faixas de pessoal ocupado

Faixa de Pessoal ocupado (indicador de porte)	Quantidade de estabelecimentos		Salários e outras remunerações (Mil Reais)		Quantidade de Pessoal ocupado	
	Qtde	(%)	R\$	(%)	Qtde	(%)
0 a 4	3.238.263	71,13	38.754.684,00	3,94	6.119.942	15,22
5 a 9	722.591	15,87	59.959.919,00	6,10	4.685.354	11,65
10 a 19	356.575	7,83	77.635.442,00	7,90	4.723.709	11,74
20 a 29	94.365	2,07	43.346.359,00	4,41	2.242.195	5,57
30 a 49	65.565	1,44	52.932.803,00	5,39	2.465.939	6,13
50 a 99	41.354	0,91	70.262.630,00	7,15	2.825.510	7,03
100 a 249	20.842	0,46	93.983.069,00	9,57	3.151.372	7,84
250 a 499	6.752	0,15	77.812.690,00	7,92	2.338.349	5,81
500 ou mais	6.124	0,13	467.711.654,00	47,61	11.667.535	29,01
Total	4.552.431	100,00	982.399.249,00	100,00	40.219.905	100,00

Fonte: Elaborado pelo autor, com dados extraídos de (IBGE, 2015).

1.6 Matemática utilizada pelo varejo

A matemática sempre esteve a frente como ferramenta auxiliar na gestão dos negócios. De um aspecto puramente prático como auxiliar para cálculo de estoques, fluxo de caixa, lotes mínimos, ela passa a tomar um aspecto mais estratégico, como a predição de receitas futuras e vendas.

Da idade antiga existem vários exemplos de uso da matemática em operações comerciais

- a) Os antigos incas usavam os quipos, bancos de memória feitos de cordas e nós utilizadas para registrar números; (PICKOVER, 2011)
- b) O papiro de Rhind é considerado como a fonte de informações mais importante conhecida relativa aos matemáticos do antigo Egito; (PICKOVER, 2011)
- c) O primeiro registro das equações de 2º grau foi feita pelos babilônios, que tinham uma álgebra bem desenvolvida e resolviam equações de 2º grau por métodos semelhantes aos atuais; (MILIES; OUTROS, 2019)
- d) Tales de Mileto mediu a distância de um navio a praia, usando, provavelmente, semelhança de triângulos. (MILIES; OUTROS, 2019)

Este processo continuou durante a Idade Média e o Renascimento, quando o crescimento das atividades comerciais deram um grande impulso no uso da matemática como ferramenta indispensável aos negócios, em especial para resolver problemas relacionados à calculo de juros, probabilidades de eventos, distancias, fracionamento de cargas e rotas.

Nos tempos modernos foram desenvolvidos métodos para solucionar problemas complexos que envolvem mais de uma variável, séries temporais, modelos de previsão de comportamento de dados amostrais, entre outros. Neste trabalho optou-se pela utilização da Estatística Descritiva em conjunto com diversos algoritmos de aprendizado de máquina, cujas principais características são:

- a) Estatística Descritiva: É o ramo da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados. Algumas das medidas normalmente usadas para descrever um conjunto de dados são as medidas de tendência central (média, mediana, moda) e as medidas de variabilidade (desvio padrão, variância, valor máximo, valor mínimo, obliquidade, curtose). (NETO, 1977)
- b) Regressão linear: É o mais simples dos modelos de regressão, baseia-se no método dos minimos quadrados também conhecido como OLS (*Ordinal Least Square*), possui um largo uso devido a sua simplicidade de formulação e processamento. (NETO, 1977)
- c) *Neural network*: Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes, que adquirem conhecimento através da experiência (FACELI; OUTROS, 2017). Elas

são compostas por várias unidades de processamento e geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, recebidos pelas suas conexões. A operação de uma unidade de processamento, proposta por McCulloch e Pitts em 1943, pode ser resumida da seguinte maneira: (CARVALHO, 2009)

1. Sinais são apresentados à entrada;
2. Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade;
3. É feita a soma ponderada dos sinais produzindo um nível de atividade;
4. Se este nível de atividade exceder um certo limite (*threshold*) a unidade produz uma determinada resposta de saída.

A primeira Rede Neural Artificial a ser implementada foi a rede perceptron, o comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede (FACELI; OUTROS, 2017).

- d) *Support Vector Machine* (SVM): As Máquinas de Vetores de Suporte (SVMs) constituem uma técnica cujos resultados da aplicação são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos, como as Redes Neurais Artificiais. As SVMs são embasadas pela teoria de aprendizado estatístico, em que estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. (FACELI; OUTROS, 2017)
- e) *Multi-Layer Perceptron* (MLP): O perceptron de múltiplas camadas (multi-layer perceptron) é uma rede neural artificial do tipo perceptron com, pelo menos, uma camada intermediária, ou seja trata-se de uma generalização do perceptron simples descrito anteriormente. (CASTRO; ZUBEN, 2003).

Uma rede MLP típica possui três características principais: (CASTRO; ZUBEN, 2003)

1. Os neurônios das camadas intermediárias (e de saída) possuem uma função de ativação não-linear do tipo sigmoidal (p. ex. função logística ou tangente hiperbólica).
 2. A rede possui uma ou mais camadas intermediárias.
 3. A rede possui altos graus de conectividade.
- f) *Random forest*: Floresta Aleatória (Random Forest) é um algoritmo de aprendizagem de máquina que cria várias árvores de decisão e as combina para obter uma predição

com acurácia. Sendo flexível e fácil de usar, pode ser utilizado para tarefas de classificação e também de regressão. (SILVA, 2018)

- g) k-Nearest Neighbour Regression (KNN): KNN é um classificador onde o aprendizado é baseado na analogia, ou seja "Objetos relacionados ao mesmo conceito são semelhantes entre si"(FACELI; OUTROS, 2017). Ele utiliza medidas de distancia para classificar um novo objeto com base nos exemplos do conjunto de treinamento que são próximos a ele. (FACELI; OUTROS, 2017)

As necessidades operacionais e administrativas necessitam de diferentes modelos de previsão. No aspecto financeiro os modelos mais utilizados são os de previsão de faturamento (valor de vendas) gerando previsões sobre o comportamento das vendas em diferentes horizontes (imediato, curto prazo, médio prazo, longo prazo) e em diferentes dimensões de tempo (horária, diária, mensal, etc.), permitindo ao responsável efetuar um planejamento financeiro.

Os modelos de previsão de demanda (quantidade de unidades vendidas) geram previsões que auxiliam na formulação de: lotes de compra, identificação de ruptura na gondola, detecção de perdas/desvios nos estoques. A criação de modelos de predição destas variáveis, em diferentes dimensões de tempo (hora, dia, mês), possibilita o seu uso na administração financeira e operacional.

Nas últimas décadas o ambiente regulatório e tecnológico em que atuam as empresas vem evoluindo de maneira consistente em direção a uma maior globalização da economia. A consequente abertura de mercados e maior concorrência obriga as empresas a encontrarem formas de se adaptar aos novos tempos, e o ganho de produtividade das empresas é a arma mais poderosa para atrair e manter a clientela, com melhores produtos e serviços a custos e preços menores. (BADIN, 1997)

OBJETIVOS

O objetivo do presente trabalho é apresentar um conjunto de ferramentas (algoritmos e métodos) que, aplicados aos dados de venda de um supermercado de pequeno porte, permitam ao administrador ter ganhos de produtividade na operação do estabelecimento.

A expectativa é que este ferramental matemático, além de auxiliar na gestão do pequeno varejista, seja uma vitrine para as principais disciplinas ofertadas pelo MECAI, demonstrando a sintonia de uma mestrado profissional com o rigor acadêmico e as necessidades reais da indústria.

2.1 Objetivos específicos

Construir um roteiro de processos e os respectivos algoritmos a serem disponibilizados, utilizando os dados de vendas de um pequeno supermercado referido como SuperX e localizado em uma região periférica da cidade de Guarulhos/SP, a partir das seguintes atividades:

1. Definir uma metodologia de aquisição dos dados adequada ao porte e condições operacionais do supermercado, com base nas melhores praticas identificadas na literatura.
2. Aplicar um processo de engenharia de dados sobre as informações disponibilizadas, compreendendo as etapas de ingestão, tratamento e disponibilização dos dados para uso em processos de matemáticos e computacionais.
3. Caracterizar os dados utilizando-se da estatística descritiva como ferramenta para identificação de padrões e desvios.
4. Gerar previsões de faturamento (valor de venda), que auxiliem e atividades de planejamento e controle financeiro do SuperX:
5. Gerar previsões de demanda (quantidade de unidades vendidas) de produtos que auxiliem na operação do SuperX:

6. Gerar previsões de preço (valor unitário de venda) de produtos que auxiliem na operação do SuperX:

2.2 Estrutura do trabalho

O presente trabalho está estruturado da seguinte forma:

O primeiro capítulo a título de introdução, tem por finalidade contextualizar o tema do projeto, em suas duas vertentes: varejo e matemática.

O segundo capítulo descreve os objetivos específicos e a estrutura do trabalho.

O terceiro capítulo trata do processo de aquisição de dados a serem utilizados.

O quarto capítulo descreve as etapas de engenharia de dados utilizadas.

O quinto capítulo demonstra o uso da estatística descritiva para compreensão e categorização dos dados.

No sexto capítulo os modelos de aprendizado de máquina são aplicados aos dados do projeto com o objetivo de se obter previsões de faturamento, demanda e preços.

No sétimo capítulo temos as conclusões.

No oitavo capítulo é apresentado o referencial bibliográfico utilizado na elaboração deste projeto.

METODOLOGIA DE AQUISIÇÃO DOS DADOS

3.1 SuperX

O supermercado chamado de SuperX para efeito de anonimato, possui 6 anos de existência, 2 terminais de pontos de venda (PDV) para atendimento aos clientes, 10 funcionários e todas as seções padrão: Padaria, Açougue, Laticínios, Frios/Congelados, Cereais. Por ter pequeno porte e uma reduzida oferta de itens, o SuperX também pode ser enquadrado como loja de conveniência.

A construção do roteiro de processos se iniciou com entrevistas junto ao proprietário e o gerente responsável pela operação do empreendimento, onde foi possível identificar problemas comuns a supermercados de sua categoria:

- a) Perdas: Furtos praticados pelos clientes, desvios praticados por funcionários e vencimento da validade dos produtos;
- b) Abastecimento: Não possui lotes mínimos para direcionar as compras, efetua compras diárias no atacarejo da região, *HortiFrutis* tem abastecimento irregular, refém de distribuidoras de bebidas (Coca-Cola, AmBev);
- c) Precificação: Trabalha com markup para definição do preço de venda, porém os produtos de maior giro (cervejas, refrigerante, açougue, cereais, padaria) possuem os menores markups devido a concorrência;
- d) Estoques: Limitações de espaço para armazenamento, itens de baixo giro ocupando espaço no mostruário, falta de produtos nas gôndolas provocando perda de vendas, compras de oportunidade por promoção/baixo preço do fornecedor acima do adequado as previsões de venda;

- e) Clientes: Desconhecimento do comportamento médio dos clientes em relação aos hábitos de compra.

Durante as entrevistas com os gestores também foi identificado que no período analisado ele passou por momentos de descontinuidade administrativa, em que houve a troca de proprietário e interferências externas que ocasionaram uma menor atenção gerencial no negócio.

O SuperX foi inaugurado em 1 de janeiro de 2013 pelo proprietário 1 e os meses iniciais foram de grande esforço na viabilização do empreendimento. Após um período de operação normal sem grandes esforços, o proprietário resolveu abrir novo negócio em outro estado, negligenciando a administração do SuperX e desviando recursos financeiros para o novo empreendimento.

Como resultado ocorreram queda nas vendas e o negócio foi vendido em 17 de novembro de 2015 para o proprietário 2. A semelhança do anterior no período inicial efetuou esforços para a recuperação do negócio e após um período de operação normal, adquiriu um novo supermercado na mesma cidade com o objetivo de ter ganhos de escala em compras, porém houve descontinuidade administrativa e o faturamento regrediu drasticamente. A situação descrita está resumida na Tabela 2:

Tabela 2 – Evolução administrativa do SuperX

Período evento	Evento	Proprietário	Eficiência administrativa
01-01-2013 a 31-03-2013	Início de atividades e busca de resultados	1	Alta
01-04-2013 a 16-08-2015	Manutenção das atividades	1	Média
17-08-2015 a 16-11-2015	Compra de estabelecimento em outro estado	1	baixa
17-11-2015 a 16-02-2016	Recuperação das atividades	2	Alta
17-02-2016 a 31-07-2017	Manutenção das atividades	2	Média
01-08-2017 a 30-09-2017	Compra de estabelecimento na mesma cidade	2	baixa

Fonte: Elaborado pelo autor, com base nas informações obtidas durante as entrevistas.

O proprietário 1 e a esposa cuidavam pessoalmente da administração do empreendimento, enquanto o proprietário 2 contratou um gerente para esta atividade, demonstrando perfis diferentes:

- a) **Proprietário 1** - Comerciante tradicional do pequeno varejo que sobrevive dos resultados do empreendimento;
- b) **Proprietário 2** - Executivo de empresas do segmento de varejo que investiu na aquisição de um empreendimento.

Esta diferença fica evidente na operação do software de Planejamento de Recursos Empresariais (ERP). O proprietário 1 utilizou o mínimo das funções, basicamente o cadastro dos produtos e a emissão do cupom fiscal, priorizando o menor trabalho administrativo e o controle gerencial da base do "olho do dono". O proprietário 2 tinha um enfoque administrativo mais formal, estabelecendo controles de compras, contas a pagar e intentou controlar os estoques pelo sistema, não atingido devido à carga de trabalho necessária.

Observou-se também uma divergência nos enfoques. O proprietário atual se mostrava mais preocupado em conseguir melhorar os resultados financeiros, e o gerente inclinado a buscar melhores resultados operacionais, não necessariamente alinhado ao melhor resultado financeiro. Outro ponto de destaque, o SuperX possui um software de gestão, porém nem todas as suas funcionalidades são utilizadas: o controle de estoque é utilizado apenas para cadastro dos itens e registro do preço de venda (lista de preços), pois é considerado muito trabalhoso registrar todas as compras item a item para ter uma gestão efetiva dos estoques. Não existe uma política de validação dos dados cadastrais, com itens erroneamente classificados em relação a seção do mesmo, provocando distorções na apresentação dos resultados por seção.

3.2 Definição de fases e processos

Os processos iniciais de KDD (*Knowledge discovery in databases*) são os mesmos adotados por (ARAÚJO, 2009) que referenciou outros autores (NETO; DINIZ, 2000; HAN; KAMBER, 2001; CORTES; PORCARO, 2002) em seu trabalho voltado para a aplicação de redes neurais em predição de vendas de supermercados.

3.2.1 Recepção

Corresponde ao estágio inicial do processo. Os dados são disponibilizados pelas fontes de origem e recebidos em seu formato original (documentos em papel, mídias eletrônicas ou interfaces com meios de comunicação eletrônica). Após o processo de leitura dos dados brutos eles estão aptos às demais etapas do processamento.(ARAÚJO, 2009)

3.2.2 Limpeza dos dados

O trabalho de limpeza dos dados ocorre no preenchimento, na padronização e correção dos dados inconsistentes. Alguns métodos de limpeza podem ser aplicados no início da etapa de KDD mas também em etapas posteriores.(ARAÚJO, 2009)

3.2.3 Valores ausentes

Na mineração de dados pode-se perceber a ausência de valores não dispostos nas bases. Existem alguns métodos que propõem soluções para este caso: (ARAÚJO, 2009)

- a) Ignorar o registro: Técnica pouco funcional, pois se torna praticável apenas quando a tupla contiver vários valores ausentes;
- b) Valor constante para preencher dados: Não muito recomendado, pode mascarar resultados dependendo da variância do atributo ou de sua importância para o processo;
- c) Valor médio de mesma classe a qual a tupla pertença: Utilizado se o valor do atributo é numérico e quando seu significado é passível de atribuição a um valor médio.

O estudo de classificação de um produto pode determinar um valor médio para um determinado valor ausente. Porém os valores fora de padrão podem influenciar consideravelmente o resultado do valor médio;

- d) Valores estatísticos para preencher dados: Pode-se utilizar técnicas de regressão ou ferramentas de inferências, tais como um formalismo Bayesiano ou indução por árvores de decisão. Método mais confiável, pois se utiliza de modelos matemáticos e considera outras informações para prever valores ausentes, dando mais confiabilidade na preservação do relacionamento entre o atributo estimado e os utilizados no processo de estimação.

3.2.4 Valores fora de padrão

São atributos que apresentam desvios acentuados ou que foram cadastrados de forma incorreta. Tópico como lucro, rendimento ou faturamento, por exemplo, estão sujeitos a este tipo de erro. Existem algumas técnicas para correção desses valores: ([ARAUJO, 2009](#))

- a) Binning: Consiste em ordenar os valores, agrupá-los, e aplicar uma medida para ajuste dos valores em cada grupo (média aritmética, mediana, valor limite), e então substituir os valores pelos calculados;
- b) Agrupamento: Valores fora do padrão podem ser detectados quando informações similares são dispostas em clusters ou grupos. Eles podem ser excluídos, mas para algumas técnicas de mineração eles também podem ser interessantes. Se o cliente quer identificar fraudes em cartões de crédito, por exemplo, os dados fora do padrão são essenciais;
- c) Inspeção humana e computador: As vezes o próprio programador pode fazer algumas medições, identificar os valores e excluí-los;
- d) Regressão: Dados podem ser ajustados por funções de regressão, à partir de duas variáveis num gráfico, desde que uma possa ser predita por outra.

3.2.5 Dados inconsistentes

São dados em não conformidade dentro de um mesmo sistema. Existem várias formas de ocorrer inconsistências num banco de dados. ([ARAUJO, 2009](#))

- a) Digitação incorreta: Quando o operador no sistema de origem atribui valores indevidos para um campo, por exemplo no campo destinado ao número do telefone ele preenche com informações de contato/recado;
- b) Erros oriundos de evolução: Utilizar informações antigas de clientes onde os números de telefones celulares ainda estão com 8 dígitos, por exemplo;

- c) Redundância de dados: Mais de uma origem para informação, normalmente temos atributos idênticos com diferentes grafias. Por exemplo ESTADO e UF representam a mesma informação e um deles deve ser descartado em processos de validação.

3.2.6 Integração de dados

Dados podem ser integrados de fontes diversas, como banco de dados, arquivos textos, *flat file*, entre outros. Este processo se assemelha ao de construção de um *Data Warehouse* (DW), construindo uma nova base consistente com os registros reunidos através das outras fontes. Três pontos devem ser considerados: (ARAÚJO, 2009)

- a) Integração de sistemas internos: Os mesmos valores semânticos podem estar incluídos em diversos esquemas com nomes e atributos diferentes, um típico caso de identificação de entidades. Em bases operacionais e em DW's os metadados minimizam esse problema;
- b) Dados redundantes: Um atributo pode ser redundante se o mesmo for derivado de outra tabela. Inconsistências em atributos ou dimensões podem ser causa de redundância em conjuntos de dados. Uma forma de tratar este problema é a utilização de análise de correlação. Esta técnica consiste em verificar o quanto dois atributos são correlatos. Outra forma de se ter atributos redundantes é tendo duas tuplas idênticas cadastradas na base de dados;
- c) Detecção e resolução de valores conflitantes: Um enorme desafio na integração dos dados está na diferença de valores que os dados podem apresentar nas diversas fontes de dados que os mesmos provêm. Em bases de dados reais, tabelas podem diferir em seus valores dependendo da localização geográfica. Por exemplo, determinados produtos podem ter seus preços variados, por conta de taxa de imposto atribuído sobre ele, de acordo com a cidade, o estado ou país em que se encontra a base.

Há outros fatores que inferem na redundância e na inconsistência dos dados, dificultando ainda mais o processo de integração que são a existência de campos fixos e variáveis e/ou diferentes formatos utilizados para o armazenamento em banco de dados relacionais, não relacionais e sistemas de arquivos dependentes do sistema operacional. Se forem tomados cuidados com as formatações dos dados e verificações de dados redundantes e inconsistentes, a integração será realizada de forma muito mais agradável, dando consistência ao processo e maior agilidade nos passos seguintes. (ARAÚJO, 2009)

3.2.7 Seleção dos dados

O método de Mineração de Dados (MD) exige que os atributos relevantes sejam escolhidos de forma a responder às perguntas que o cliente quer saber, e por isso é importante a participação de pessoas ligadas ao negócio em que as tarefas de MD serão aplicadas. Se

o cliente quer saber se um determinado produto sai em sequência de outro, seleciona-se os atributos relacionados com as transações comerciais do estabelecimento e aplica-se uma técnica de regra de sequência, definindo que diante da saída de um determinado produto, algum tempo depois outro produto associado a este será vendido também. Essa técnica pode ser utilizada na análise de compras de rádios automotivos, por exemplo. Algum tempo após a venda de um rádio, autos-falantes podem ser vendidos também, ajudando na estratégia da empresa de combinar a venda destes produtos. (ARAÚJO, 2009)

3.2.8 Transformação dos dados

Nesta etapa é enriquecedor mudar a semântica dos dados ou atributos e adaptá-los à nova forma da aplicação que realizará a MD. As principais regras de transformação são: (ARAÚJO, 2009)

- a) Agregação: Agrega e sumariza os dados em uma tabela de vendas diárias. Estes atributos são agregados em vendas semanais, mensais e anuais;
- b) Aplainamento: Retira dados ruidosos. Utilizam técnicas de agrupamento, binning e regressão;
- c) Generalização: Dados podem ser alterados para um contexto mais abstrato. O atributo idade pode ser alterado para faixa etária;
- d) Construção de atributos: Atributos novos são construídos de acordo com informações existentes.
- e) Redução de dados; dividido em:
 - Agregações, Sumariza vários registros de uma informação em único registro normalmente utilizando funções matemáticas para os dados numéricos (soma, média, etc.);
 - Redução dimensional – Elimina atributos irrelevantes à técnica utilizada;
 - Compressão dos dados – Utiliza de codificação para reduzir o conjunto de dados;

3.3 Origem dos dados

Foram fornecidos pelo SuperX vários arquivos em formato CSV, contendo as transações do período de jan-2013 a dez-2015. A análise inicial mostrou existir um erro no processo de geração dos arquivos, pois todos ficaram limitados a 65536 registros, ou seja não existem dados disponíveis para todo o período informado. Após análise foi identificado que poderiam ser geradas novas variáveis com base nas informações originais, capturando características como: período da venda, semestre, semana do ano, etc.

Em Janeiro de 2018 o SuperX forneceu novos dados em arquivos do formato .GDB, no formato proprietário do banco de dados Firebird, que é a solução de banco de dados utilizada pelo software de gestão do SuperX para armazenar todas as informações referentes as operações realizadas.

Por questões de economia com licenças de uso de software e possibilidade de utilizar hardware de baixo poder de processamento, muitas empresas desenvolvedoras optam por utilizar *software Open Source* em seus projetos, como forma de atingir um público com restrições de orçamento, o caso de supermercados de pequeno porte.

Um componente comum nestas soluções é o Firebird, um Sistema Gerenciador de Base de Dados (SGBD) *open source* originário do código fonte do Interbase que foi disponibilizado pela Borland em 2000 como Software Livre.((REEVES, 2001))

Com base no aprendizado sobre os dados obtidos com os arquivos em formato .CSV não ocorreram maiores dificuldades no processamento destes arquivos, e foi identificado que este banco continha a totalidade das informações no período de estudo.

3.4 Processamento de dados

Após a caracterização das fases e processos usuais em atividades de aquisição de dados para mineração, foram definidos os seguintes passos a serem aplicados: (ARAÚJO, 2009)

- a) Leitura dos dados: Os arquivos recebidos foram armazenados em um diretório padrão, e então lidos a partir de uma rotina escrita em linguagem R, a validação física dos arquivos com a observação da inexistência de erros e a geração de um quadro resumo dos dados lidos são as tarefas executadas nesta etapa;
- b) Limpeza dos dados: Dadas as características de origem dos dados a serem lidos e posteriormente processados, foram adotados as seguintes regras:
 1. Valores ausentes: Entende-se que a ausência de algum valor inutiliza o registro, pois os dados de venda sempre devem ser entregues com todos os seus valores, e a ausência indica falha no processo de geração dos valores e/ou do processo de armazenamento dos mesmos. Desta forma quaisquer registros com dados faltantes (nulos) devem ser descartados;
 2. Valores fora do padrão: Foram removidos os registros cujos valores de venda eram superiores a R\$ 500,00 (quinhentos reais).
- c) Integração dos dados: Para cada arquivo fornecido é executado o processo de leitura e limpeza dos dados, e em seguida os dados resultantes são adicionados ao arquivo SaidaTotal.csv. Após o processamento de todos os arquivos para eliminar dados

redundantes, é executada uma etapa adicional para exclusão de eventuais registros duplicados no arquivo SaidaTotal.csv;

- d) Seleção dos dados: Apenas as variáveis diretamente relacionadas ao produto vendido são de interesse para este estudo. Desta forma variáveis que representam informações fiscais, financeiras e de controle por exemplo são excluídas;
- e) Transformação de dados: Em função dos dados recebidos é possível a criação de novos dados que expõem de forma mais clara as propriedades da transação (registro da venda) , estes novos dados podem ter relevância na execução de modelos preditivos em substituição ou reforço as variáveis originais. Um exemplo é extrair a informação de fim-de-semana da data de venda e trata-la como uma variável adicional.
- f) Disponibilização dos dados: Após todas as etapas anteriores, obtêm-se um arquivo final contendo todas as transações que efetivamente podem ser utilizadas para a elaboração de análises estatísticas e a construção de modelos preditivos.

ENGENHARIA DE DADOS

Utilizando o processo de engenharia de dados, foram codificadas diversas rotinas em linguagem R, que executaram as funções especificadas na Figura 2. O objetivo é disponibilizar os dados em formato pertinente ao seu uso nos algoritmos de uso comum em ciência de dados (algoritmos estatísticos, matemáticos e de aprendizado de máquina).

Figura 2 – Fluxograma do processo de engenharia de dados

Fonte: Elaborado pelo autor

A escolha da linguagem R, deveu-se ao seu amplo uso nas comunidades acadêmicas e empresarial na solução de problemas envolvendo análise de dados.

"R é uma linguagem e ambiente para computação estatística e gráfica, Ela é um projeto GNU similar a linguagem S que foi desenvolvida nos laboratórios *Bell* (Antiga *AT&T*, atual *Lucent Technologies*). R providencia uma grande quantidade de métodos gráficos e estatísticos (modelos lineares e não lineares, testes estatísticos clássicos, análise de series temporais, classificação, clusterização entre outros) e é altamente extensível. O código fonte do R é disponibilizado como software livre

sobre os termos da *Free Software Foundation* sob uma licença *General Public License* (GNU)"([R-PROJECT, 1997](#))

A existência de inúmeras bibliotecas especializadas nas atividades de interesse para este trabalho (mineração de dados, *machine learning*, visualização) foram de fundamental importância na finalização das atividades.

4.1 Aquisição dos dados

A aquisição dos dados é uma das atividades cruciais em projetos que envolvam dados de fontes externas, e foi dividida em 3 etapas:

4.1.1 Infraestrutura

Para o processamento dos dados, incluídas as atividades de codificação e depuração foram utilizados os seguintes equipamentos: desktop com processador Core 2 Duo, 8 GB memória RAM e sistema operacional Windows 10; notebook com processador Core I7, 12 GB memória RAM e sistema operacional Windows 10;

À exceção do sistema operacional, foram utilizados apenas *softwares* livres em todas as etapas do processo: linguagem R + bibliotecas especializadas; IDE R Studio; banco de dados *Firebase* com *drives* de conexão; ferramenta de consulta *Squirrel*.

4.1.2 Hacking

O conjunto de arquivos fornecido pelo proprietário do SuperX em formato utilizado pelo banco de dados Firebird para armazenagem de dados (extensão .gbd). Não possuía indicações de acesso (usuário, senha) a serem utilizados pois o fornecedor do software de PDV utilizado pelo SuperX não forneceu estes dados. Após consulta na *internet* foi identificado que os dados de acesso padrão para o Firebird são SYSBDA e *masterkey* para login e senha, respectivamente. Eles se mostraram eficazes, e embora representem uma vulnerabilidade que deveria ser corrigido pelo produtor do software, a partir destas credenciais foi possível a execução do Squirrel para validação do acesso aos dados, e a identificação das tabelas e colunas de interesse para a montagem da consulta Structured Query Language (SQL).

4.1.3 Leitura de dados

Codificação de rotina para ler os dados existentes na Firebase, composta de comandos para: criar conexão com o banco de dados; listar todas as tabelas do banco de dados; definir a consulta SQL; executar a consulta SQL e gravar os dados lidos em formato .CSV. Ao final do

processo foram extraídos 2.242.917 registros com 83 colunas, posteriormente gravados em disco local, resultando no arquivo result_mov.csv com um tamanho de 921.199.778 bytes.

4.2 Seleção dos dados

Do conjunto de registros obtidos foi necessário filtrar e eliminar aqueles que não são de interesse. Os parâmetros que justificam a eliminação dos registros são:

- a) Não são itens do cupom de venda (todos os registros com atributo TIPOREGISTRO $\neq$ de "VI");
- b) Itens cancelados (todos os registros com o atributo ITEMCANCELADO=="S");
- c) Itens com valor total do item = ZERO (todo registro com atributo VALORTOTALITEM==0).

Após o processo de seleção a amostra foi reduzida a 2.227.786 registros.

4.3 Seleção de atributos

A seleção de atributos consiste na identificação das características que não são de interesse para o estudo, como os relacionados às informações fiscais, ao controle do processo de venda e à configuração do sistema mas que não são utilizados pela gerência (dados operacionais das promoções, estoques, parcelamento venda, etc.). Os atributos e suas indicações de dimensionalidade podem ser observados na Tabela 3:

Também foram eliminados atributos que apresentam ausência de dados e todos aqueles que se encaixavam em algum dos critérios abaixo.

- a) Quantidade de valores únicos = 1, valores constantes não tem interesse em modelos de predição;
- b) Quantidade de valores nulos > 0 , existem processos que não permitem a existência de valores nulos, como a análise de componentes principais;
- c) Atributo contendo informações fiscais, não existe a preocupação de se cadastrar corretamente essas informações fiscais;
- d) Atributo contendo informações financeiras, as de cunho financeiro não são alimentadas pelo operador de caixa (pagamento parcelado, com cheque, etc.);
- e) Atributo contendo informações operacionais, como supervisor, caixa, que são reutilizadas em caso de troca, não permitindo a identificação correta.

Desta forma apenas os atributos diretamente relacionados ao processo de venda foram selecionados, como enumerado na Tabela 4:

Tabela 3 – Lista de atributos disponibilizados pelo SuperX

	NomeColuna	QtdeValoresUnicos	QtdeValoresNulos	PercValoresNulos
1	codigoproduto	9106	0	0
2	descricaoCompleta	9106	0	0
3	codigogruppo	105	0	0
4	grupo	105	0	0
5	codigosubgrupo	20	0	0
6	subgrupo	344	0	0
7	codigodepartamento	2	0	0
8	departamento	2	0	0
9	codigosecao	21	0	0
10	secao	21	0	0
11	codigo	2227786	0	0
12	numerocupomfiscal	870360	0	0
13	numerocupomnfiscal	870360	0	0
14	caixa	2	0	0
15	serialecf	4	0	0
16	datacupom	1728	0	0
17	hora	54948	0	0
18	operador	11	0	0
19	vendedor	1	0	0
20	loja	1	0	0
21	finalizadora	1	0	0
22	codigocliente	1	0	0
23	parcelas	1	0	0
24	numeroitem	234	0	0
25	numerovolume	232	0	0
26	numerocuponscancelados	1	0	0
27	numeropedido	1	0	0
28	tipodocumento	1	0	0
29	tiporegistro	1	0	0
30	cpf	1	0	0
31	nome	1	0	0
32	telefone	1	0	0
33	banco	1	0	0
34	agencia	1	0	0
35	contacorrente	1	0	0
36	numerocheque	2	0	0
37	sequencia	196	0	0
38	codigoproduto	9106	0	0
39	codigobarras	85759	0	0
40	descricaooproduto	5736	0	0
41	tipoproduto	1	0	0
42	formula	1	0	0
43	quantidade	5766	0	0
44	valorunitario	628	0	0
45	valortotalitem	6097	0	0
46	tributacaopdv	2	0	0
47	unidadevenda	9	0	0
48	ecf	2	0	0
49	codigotributacao	5	0	0
50	itemcancelado	1	0	0
51	cupomcancelado	2	0	0
52	valorsubtotal	16752	0	0
53	valortotalcupom	1	0	0
54	valorrecebido	1	0	0
55	valortroco	1	0	0
56	valordesconto	3	0	0
57	valor acrescimo	1	0	0
58	tipovenda	1	0	0
59	juroschegue	1	0	0
60	jurosticket	1	0	0
61	fundocaixa	1	0	0
62	contravale	1	0	0
63	milhagem	1	0	0
64	descontocupom	8	0	0
65	precoatacado	5595	0	0
66	descatacado	1	0	0
67	minimoatacado	2	0	0
68	numeroitenscancelados	1	0	0
69	valoritenscancelados	1	0	0
70	valorcuponscancelados	1	0	0
71	codigoconvenio	1	0	0
72	codigoconveniado	1	0	0
73	vencimentocheque	1	0	0
74	supervisor	11	0	0
75	transferido	2	0	0
76	descricaoatpg	1	0	0
77	modalidadeatpg	1	0	0
78	tipopagamento	1	0	0
79	versao	5	0	0
80	custo	1	2227786	100
81	cfcanc	1	2227786	100
82	nseriesat	1	2227786	100
83	chavecfe	1	2227786	100

Fonte: Elaborado pelo autor

4.4 Limpeza dos dados selecionados

Identificação e eliminação das anomalias existentes nos registros de vendas, ou seja, todos os itens cujo valor total sejam superiores a R\$ 500,00 (quinhentos reais).

A regra foi definida junto com o gestor do SuperX, que informou que eventuais vendas acima deste valor são, na maioria das vezes, fruto de vendas esporádicas por atacado (para bares, restaurantes), vendas para eventos como churrascos e aniversários, ou lançamentos para ajustes no caixa, não representando de forma alguma um movimento normal de venda.

Também foram identificados ocorrências de cadastro incorreto dos produtos. A Figura 3, mostra exemplo que o produto de código 21132 (prendedor de roupa) foi atribuído ao departamento de perecíveis, seção de laticínios, no grupo de iogurtes, quando o correto era departamento de não perecíveis, seção de utilidades, no grupo domésticas.

Tabela 4 – Lista de atributos selecionados para o SuperX

	Nome do atributo
1	codigoproduto
2	descricao completa
3	codigogrupo
4	grupo
5	codigosubgrupo
6	subgrupo
7	codigodepartamento
8	departamento
9	codigosecao
10	secao
11	codigo
12	numerocupomfiscal
13	caixa
14	datacupom
15	hora
16	operador
17	numeroitem
18	sequencia
19	quantidade
20	valorunitario
21	valortotalitem
22	unidadevenda
23	codigotributacao
24	supervisor

Fonte: Elaborado pelo autor

Para a mitigação do problema da qualidade do cadastro de produtos, foram verificadas duas alternativas.

- a) Classificar os produtos por ordem descendente de faturamento, agrupando os produtos menos representativos. Tem o efeito colateral de inutilizar as informações existente sobre diversos atributos: hora, operador, número cupom. Foi criado um programa para a consolidação da codificação dos produtos de acordo com a seguinte regra: os produtos que representam 80% das vendas mantiveram o seu código original, o restante foi agrupado em pacotes que representavam um valor percentual do faturamento maior que o do último produto. Como resultado foi possível diminuir de 9106 para apenas 2086 itens cadastrados;
- b) Reclassificação manual dos produtos, categorizando-os corretamente quanto a seus atributos: departamento, seção, grupo e incluir manualmente as informações referentes a subgrupo e marca. Alternativa extremamente trabalhosa, porém enriquece em muito os dados. Para manter a reprodutibilidade do processo foi codificada rotina em R para efetuar todas as alterações detectadas como necessárias na massa de dados. Foram alteradas descrições, classificação (departamento, seção, grupo) e criados novos atributos (marca, subgrupo).

Para preservar o maior número de informações possível e também para eliminar as possíveis classificações incorretas, optou-se pela segunda alternativa. Embora mais trabalhosa -

Figura 3 – Produtos cadastrados erroneamente

Fonte: Elaborado pelo autor

o processo de reclassificação manual dos produtos levou mais de 70 dias-, ela preserva a riqueza de detalhes dos dados iniciais. O resultado final está sintetizado na Figura 4:

Figura 4 – Reclassificação dos produtos

	Tipo classificação	Original	Reclassificado
1	Departamento	2	3
2	Secao	21	14
3	Grupo	105	102
4	Subgrupo	412	442

Fonte: Elaborado pelo autor

Após o processo de limpeza foram efetuadas validações sobre a massa de dados em busca dos possíveis problemas: existência de valores indeterminados; vendas com quantidade = 0; vendas com valor unitário = 0; e vendas com valor total = 0.

4.5 Integração dos dados

A integração de dados externos é uma etapa que pode melhorar a qualidade da análise e modelagem matemática. Foram identificadas e consideradas as seguintes origens:

- a) Entrevista com os proprietários detalhada na Tabela 1, sobre a troca de proprietário e mudanças de qualidade do acompanhamento gerencial da empresa. Para representar estes eventos foram adicionados 2 novos atributos: proprietário (assumindo os valores 1 e 2) e qualidade de gestão (assumindo os valores: forte para os primeiros 90 dias, normal durante o período intermediário e fraco para os últimos 90 dias);
- b) Feriados são um evento que provocam impactos no faturamento. Foi recuperada a tabela de feriados mantida pela Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA) em seu site da internet no endereço <http://www.anbima.com.br/feriados/feriados.asp>. Esta tabela é utilizada para indicar os feriados em que não há sensibilização das Reservas Bancárias, desta forma foi necessário acrescentar manualmente os feriados referentes ao município de operação do SuperX, o feriado estadual de 09 de julho válido no estado de São Paulo e os feriados por motivos eleitorais;
- c) Inflação é um fenômeno econômico que pode distorcer análises. Para mitigar este problema e permitir uma análise a valor constante quando necessário, foram incorporadas séries históricas de índices econômicos (INPC - Índice Nacional de Preços ao Consumidor e IPCA - Índice de Preços ao Consumidor) fornecidos pelo IBGE e calculado o valor total do item corrigido pela variação do índice no período da venda (mês de emissão do cupom fiscal) e set/2017 (fim do período analisado).

4.6 Transformação dos dados

Como etapa final foram efetuadas as seguintes transformações:

- a) Criação de atributos temporais que representam frações de um dia:
 - hora_dia No formato HH:MM:SS identifica o horário do dia em que ocorreu a venda;
 - periodo_dia Identifica o período do dia em que ocorreu a venda de acordo com a seguinte regra:
 1. HORA $\geq$ 5 e HORA $<$ 11 = 'manha',
 2. HORA $\geq$ 11 e HORA $<$ 15 = 'meio_dia',
 3. HORA $\geq$ 15 e HORA $<$ 19 = 'tarde'
 4. HORA $\geq$ 19 ou HORA $<$ 5 = 'noite'
- b) Criação de atributos temporais que representam ciclos ou múltiplos de 1 dia:
 - Dia: Valor numérico no intervalo de 01 a 31, que identifica o dia do mês em que ocorreu a venda;

- Mês: Valor numérico no intervalo de 01 a 12, que identifica o mês do ano em que ocorreu a venda;
 - Ano: Valor numérico que registra o ano que ocorreu a venda;
 - Dia_semana: Identifica o dia da semana em que ocorreu a venda;
 - Fim_semana: Identifica se a venda ocorreu em um fim de semana ou não
 - Nr_semana Identifica o número da semana no ano;
 - Trimestre: Identifica o trimestre associado a data em que ocorreu a venda;
 - Semestre: Identifica o semestre associado a data em que ocorreu a venda;
 - Estação do ano (Verão, Outono, Inverno e Primavera), associado a data em que ocorreu a venda;
 - Nr_quinzena Identifica se primeira ou segunda quinzena no mês do ano em que ocorreu a venda.
- c) Categorização das vendas por diferentes níveis de aglutinação (departamento, seção, grupo e produto), sendo criados variáveis para armazenar o valor acumulado das vendas, a classificação deste valor em ordem crescente e a representação percentual do valor acumulado pra cada um dos níveis.

Em seguida foram efetuados ajustes nos nomes das colunas geradas eliminando caracteres especiais e espaços em branco (por exemplo: "grupo_TEMPEROS / MOLHOS" virou "grupo_temperos_molhos"), e as colunas com dados alfanuméricos foram convertidas em fatores (lista de valores no R). Os dados foram armazenados no disco local em formato CSV, e em formato binário do R para permitir o reuso do mesmo em outras rotinas sem a necessidade de processamentos adicionais.

4.7 Pré-processamento

Após a etapa de transformação dos dados, o conjunto de informações está disponível para ser utilizado em diferentes algoritmos de aprendizado de máquina (*Machine learning*), porém existem transformações adicionais que são efetuadas com o objetivo de:

1. Atender as particularidades do algoritmo;
2. Melhorar a acurácia nos resultados;
3. Consolidar os dados de acordo com uma dimensão temporal e/ou hierarquia organizacional dos dados;
4. Remover atributos que não são de interesse para o estudo;
5. Remover atributos que não são adequados ao algoritmo;

4.8 Observações

Foi identificada a existência de erros originados de falhas no processo de operação do sistema, devido a:

1. Cadastramento incorreto de produtos (erros na classificação: departamento, seção, grupo);
2. Não cadastramento de produtos (no PDV o caixa não localiza o produto e utiliza o código de outro);
3. Reutilização de códigos (operadores e supervisores, quando admitidos utilizavam o mesmo código do funcionário anterior).

Estes erros foram devidamente tratados e os dados enriquecidos com novos atributos gerando um conjunto de dados compatíveis com os modelos utilizados na Ciência de Dados. Para a atividade de engenharia de dados foram codificados os programas listados na Tabela 5:

Tabela 5 – Lista de artefatos *codigofonte* – *EngenhariadeDados*

Programa	Descrição resumida
P000_ConfiguracaoGlobal_V3	Inicializa variáveis globais.
P001_LeituraDados_V4	Lê os dados obtidos no processo de Engenharia de Dados e disponibiliza para processamento pelas rotinas de análise estatística e de <i>machine learning</i> .
P010_AquisicaoDados_V4	Obtem dados existentes em banco de dados Firebird e disponibiliza para processamento.
P011_AjusteDadosProdutos_V3	Qualifica o cadastro de produtos de acordo com a interpretação do proprietário (correção manual de informações cadastrais).
P012_IntegracaoTransformacaoDados_V3	Efetua processos de integração e transformação sobre os dados adquiridos, inclusive com os dados ajustados dos produtos.
P013_AgruparProdutos_V1	Consolida as vendas dos produtos menos relevantes, diminuindo a quantidade total de produtos a serem analisados.

Fonte: Elaborado pelo autor

ESTATÍSTICA DESCRITIVA

De acordo com [Neto \(1977\)](#), a estatística descritiva é o conjunto de ferramentas e métodos estatísticos que tem por objetivo resumir e descrever as características importantes de um conjunto de dados. O insumo utilizado para a aplicação da análise univariada (estatística descritiva) foi o conjunto de dados obtidos no processo de Engenharia de Dados.

As informações e a caracterização dos dados obtidos no processo de engenharia de dados foi efetuada em etapas, variando a complexidade do agrupamento de acordo com o nível hierárquico da informação agrupada na tabela 6:

Tabela 6 – Níveis hierárquicos da informação

Dimensão Produto	Dimensão Tempo
1 - Loja	A - Ano
2 - Departamento	M - Mês
3 - Seção	D - Dia
4 - Grupo	H - Hora
5 - Produto	M - Momento (Data e hora)

Fonte: Elaborado pelo autor

O conjunto de dados possui duas hierarquias, onde uma determina os diferentes níveis de classificação dos produtos, e a outra identifica as informações referentes a variação temporal das informações, e apresentam algumas particularidades:

- Dimensão Produto: os níveis de sub-grupo e modelo não foram considerados, pois não estão devidamente definidos, visto que os dados fornecidos pelo SuperX tinham problemas de qualidade impossibilitando a sua extração durante o processo de engenharia de dados.
- Dimensão Tempo: como o conjunto de dados do estudo representa uma série temporal de observações, existe a possibilidade de derivar novos níveis a partir de operações de

agrupamento na dimensão tempo:

- semana (7 dias);
- quinzena (15 dias);
- bimestre (2 meses);
- semestre (6 meses);
- quinquênio (de 5 anos).

Para dar suporte as análises que foram efetuadas, em todas as situações foram calculados os seguintes atributos, mostrados na Tabela 7.

Tabela 7 – Atributos derivados de sumarizações e cálculos

Atributo	Expressão	Descrição
qtde_vendas	n()	Quantidade de ocorrências
qtde_unidades	sum(quantidade)	Somatória das quantidades
valor_venda	sum(valortotalitem)	Somatória do valor de venda em R\$
valor_venda_inpc	sum(valortotalitem_inpc)	Somatória do valor de venda corrigido pelo INPC
valor_venda_ipca	sum(valortotalitem_ipca)	Somatória do valor de venda corrigido pelo IPCA
qtde_vendas_perc	qtde_vendas / sum(qtde_vendas) * 100	Participação percentual do valor das vendas no mês sobre o total de vendas no período
qtde_unidades_perc	qtde_unidades / sum(qtde_unidades) * 100	Participação percentual do valor das vendas no mês sobre o total de vendas no período
valor_perc	valor_venda / sum(valor_venda) * 100	Participação percentual do valor das vendas no mês sobre o total de vendas no período

Fonte: Elaborado pelo autor

Dentro do processo de análise estatística destaca-se as medidas selecionadas para descrever o conjunto de dados utilizados.

- Medidas de tendencia central, neste trabalho será utilizado a média e a mediana;
- Medidas de variabilidade ou dispersão, neste trabalho será utilizado o desvio padrão, o valor máximo e o valor mínimo.

O desvio padrão é a medida que expressa o grau de dispersão de um conjunto de dados em relação ao seu valor médio, quanto mais próximo de 0 for o desvio padrão, mais homogêneos são os dados. O desvio padrão de uma população é calculado pela formula 5.1:

$$DP = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (5.1)$$

Onde:

DP = Desvio padrão

n = numero de casos

$\bar{x}$ = valor médio de x

x_i = variável x

Quando o cálculo do desvio padrão é executado sobre uma parte da população (amostra) são utilizadas as mesmas variáveis do cálculo do desvio padrão populacional, porém a formula utilizada é ligeiramente diferente, pois o denominador deve ser subtraído de 1 unidade quando trata-se do cálculo do desvio padrão amostral, conforme formula 5.2 abaixo:

$$DP = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (5.2)$$

5.1 Análise exploratória de dados

Foi utilizada a package SmartEDA, para a execução da análise exploratória de dados no R, obtendo os resultados apresentados na figura 5. É possível observar que o processo de engenharia de dados realizado na etapa anterior foi bastante eficiente, pois não há ocorrência de variáveis cujo valor é desconhecido.

Figura 5 – Visão geral dos dados analisados

	Descriptions	Obs
1	Sample size (Nrow)	2101510
2	No. of Variables (Ncol)	69
3	No. of Numeric Variables	39
4	No. of Factor Variables	28
5	No. of Text Variables	0
6	No. of Logical Variables	0
7	No. of Date Variables	1
8	No. of Zero variance Variables (Uniform)	1
9	%. of Variables having complete cases	100% (69)
10	%. of Variables having <50% missing cases	0% (0)
11	%. of Variables having >50% missing cases	0% (0)
12	%. of Variables having >90% missing cases	0% (0)

Fonte: Elaborado pelo autor

A estrutura das variáveis que compõem o conjunto de dados está documentada na figura 6:

Figura 6 – Estrutura dos dados

S.no	Variable Name	Variable Type	% of Missing	No. of Unique values
1	codigoproduto*	factor	0	9106
2	grupo*	factor	0	95
3	secao*	factor	0	14
4	departamento*	factor	0	3
5	codigo*	factor	0	2101510
6	datacupom**	Date	0	1728
7	numerocupomfiscal*	factor	0	870337
8	descriçãocompleta*	factor	0	9095
9	caixa*	factor	0	2
10	hora	numeric	0	54948
11	operador*	factor	0	11
12	numerositem*	factor	0	211
13	sequencia*	factor	0	153
14	quantidade	numeric	0	6885
15	valorunitario	numeric	0	627
16	valortotalitem	numeric	0	7115
17	unidadevenda*	factor	0	9
18	codigotributacao*	factor	0	5
19	supervisor*	factor	0	11
20	subgrupo*	factor	0	377
21	embalagem_padrao*	factor	0	1316
22	marca_home*	factor	0	47
23	qtde_cupom_fiscal	numeric	0	1

S.no	Variable Name	Variable Type	% of Missing	No. of Unique values
24	qtde_item_cupom_fiscal	numeric	0	18
25	qtde_produto	integer	0	18
26	feriado*	factor	0	2
27	proprietario*	factor	0	2
28	qualidade_gestao*	factor	0	3
29	indice_ipca	numeric	0	57
30	perc_ipca	numeric	0	48
31	indice_inpc	numeric	0	57
32	perc_inpc	numeric	0	50
33	valortotalitem_ipca	numeric	0	114033
34	valortotalitem_inpc	numeric	0	114109
35	hora_dia*	factor	0	17
36	periodo_dia*	factor	0	4
37	dia	integer	0	31
38	mes	numeric	0	12
39	ano	numeric	0	5
40	dia_semana*	factor	0	7
41	fim_semana	numeric	0	2
42	nr_semana	numeric	0	53
43	trimestre	numeric	0	4
44	semestre	numeric	0	2
45	estacao_ano*	factor	0	4
46	departamento_qtde_vendas	integer	0	3

S.no	Variable Name	Variable Type	% of Missing	No. of Unique values
47	departamento_qtde_unidades	numeric	0	3
48	departamento_valor_venda	numeric	0	3
49	departamento_valor_acumulado	numeric	0	3
50	departamento_ordem_faturamento*	factor	0	3
51	departamento_perc_faturamento_acumulado	numeric	0	3
52	secao_qtde_vendas	integer	0	14
53	secao_qtde_unidades	numeric	0	14
54	secao_valor_venda	numeric	0	14
55	secao_valor_acumulado	numeric	0	14
56	secao_ordem_faturamento*	factor	0	14
57	secao_perc_faturamento_acumulado	numeric	0	14
58	grupo_qtde_vendas	integer	0	95
59	grupo_qtde_unidades	numeric	0	95
60	grupo_valor_venda	numeric	0	95
61	grupo_valor_acumulado	numeric	0	95
62	grupo_ordem_faturamento*	factor	0	95
63	grupo_perc_faturamento_acumulado	numeric	0	95
64	produto_qtde_vendas	integer	0	991
65	produto_qtde_unidades	numeric	0	1229
66	produto_valor_venda	numeric	0	6617
67	produto_valor_acumulado	numeric	0	9106
68	produto_ordem_faturamento*	factor	0	9106
69	produto_perc_faturamento_acumulado	numeric	0	9106

Fonte: Elaborado pelo autor

Seguindo a classificação dos níveis de informação definidas na Tabela 4, foram efetuadas análises para cada uma das dimensões do produto, e em diferentes combinações da dimensão tempo, agregando os dados às respectivas dimensões e a visualização dos dados resultantes.

5.2 Análise dos dados

O SuperX é um empreendimento sem filiais, depósitos ou qualquer espécie de divisões administrativas, desta forma não existem agregações a serem consideradas na dimensão organizacional.

Os dados foram agregados sem considerar suas sub divisões (departamentos, seções) e foram exploradas as variações da dimensão temporal segundo dois critérios:

- Histórico: Para visualizar as informações de faturamento mês a mês e obter uma perspectiva de comportamento temporal das vendas;
- Sazonal: Para identificar padrões de comportamento das vendas que se repetem ao longo do tempo.

5.2.1 Faturamento mensal

Para obter uma visão histórica da evolução do faturamento do SuperX, os dados foram agrupados por mês e ano, e os atributos definidos na tabela 7 foram calculados gerando um novo conjunto de dados:

A figura 7 apresenta a evolução do faturamento mensal em R\$ e em moeda constante (R\$ corrigidos pela variação do INPC e IPCA respectivamente). O valor do faturamento expresso em R\$ apresenta uma leve tendencia de crescimento, porém quando corrigidos os valores pela

variação do INPC e do IPCA, verifica-se que em termos reais o valor do faturamento decresceu, um comportamento contrário ao esperado pelo proprietário do SuperX.

Figura 7 – Faturamento mensal deflacionado

Fonte: Elaborado pelo autor

No período de jan/2013 até set/2017 os índices INPC e IPCA têm variação acumulada de 33.42% e 33.78% respectivamente.

A figura 8 mostra a participação percentual dos indicadores, ou seja apresenta a participação mensal em termos percentuais das quantidades de unidades vendidas, das quantidades de vendas e do valor da venda.

Figura 8 – Indicadores de vendas - participação mensal

Fonte: Elaborado pelo autor

Observa-se um deslocamento na posição das linhas do gráfico, onde a quantidade de vendas (cupons emitidos) reduziu sua participação indicando queda, e a quantidade de unidades

vendas cresceu. Em situações onde o resultado não pode ser explicado por uma estratégia da empresa, é um indicador de queda no valor médio das vendas.

A figura 9 mede a variação percentual das quantidades de unidades vendidas, das quantidades de vendas e do valor da venda em R\$ e corrigido pelo INPC, tomando como base jan/2013 e o respectivo mês.

Figura 9 – Indicadores de vendas - variação percentual

Fonte: Elaborado pelo autor

Em consonância com o gráfico de faturamento mensal, observa-se que o faturamento em valores reais apresenta tendência declinante, porém quando medido em valores corrigidos pelo INPC a queda é bem mais brusca.

A análise dos gráficos contidos nas figuras 8 e 9 identifica que a participação percentual dos indicadores da quantidade de itens vendidos sofreu um grande aumento nos últimos meses analisados. Segundo o proprietário isto é reflexo da falta de cadastramento dos produtos no sistema de PDV, então toda vez que o produto não estava cadastrado o caixa simulava uma venda de balas com preço unitário de R\$ 0,01 na quantidade necessária para efetuar a cobrança do item.

A figura 10 mostra o resumo do faturamento com base anual, em que o volume financeiro envolvido é bastante alto. Em todo o período analisado a inflação acumulada pelo INPC é de 33.42%, e o valor das vendas cresceram apenas 9.61% em termos nominais, demonstrando as dificuldades enfrentadas pela gerência na gestão do negócio.

A queda do faturamento em valores reais e a falha no cadastramento dos produtos evidenciam a existência de problemas operacionais e administrativos na gestão do SuperX, e devem ser melhor explorados.

5.2.2 Faturamento sazonal

Segundo (SHIKIDA; MARGARIDO, 2009)

Figura 10 – Faturamento anual: Realizado X Ajustado pela inflação

	ano	qtde_vendas	qtde_unidades	valor_venda	valor_venda_inpc	valor_venda_ipca
1	2013	449838	517976.19	1575464.02	2054807.27	2058510.58
2	2014	452444	578098.56	1781811.34	2191479.57	2189368.77
3	2015	477300	601425.88	1967993.01	2220441.88	2224379.19
4	2016	431970	622087.33	1924735.39	1982204.11	1996490.17
5	2017	289958	877272.08	1329172.89	1331822.82	1337374.16

Fonte: Elaborado pelo autor

"O componente sazonal capta os padrões regulares da série de tempo, tais como mudanças de temperatura, índice pluviométrico, safra ou entressafra de produtos agropecuários, vendas da indústria, vendas do varejo, entre outros."

Para o atual conjunto de dados foram considerados como fatores sazonais a hora do dia, o dia da semana e o mês do ano.

A figura 11 demonstra o comportamento dos indicadores de venda por hora do dia. É possível identificar uma curva no formato aproximado de um M, indicando que as vendas crescem até o período de preparação de almoço por volta das 11h00 e caem abruptamente, ficando estáveis no período entre 13h00 e 15h00, quando então volta a subir no período do jantar.

Figura 11 – (%) do faturamento por hora do dia

Fonte: Elaborado pelo autor

Na primeira hora de atendimento (das 06h00 as 07h00) as vendas não são expressivas, e no período final do dia os dados das 2 últimas horas (das 20h00 as 22h00) são reflexo de

situações excepcionais (vésperas de feriados, período natalino, etc.) pois o SuperX encerra as atividades as 20h00 normalmente.

Outro ponto identificado é que no horário de almoço existe um aumento no percentual da quantidade de vendas realizadas e do percentual do valor das mercadorias vendidas, em contraste com o percentual de unidades vendidas que apresenta uma leve queda, indicando a venda de produtos de maior valor agregado, neste caso os produtos do açougue. Considerando que o pico de vendas ocorre as 19h00 e no horário de encerramento (20h00) é o momento que possui o segundo maior volume de vendas, existe um forte indicio que o horário deveria ser estendido para pelo menos até as 21h00 para atender a uma demanda existente e aumentar o faturamento do dia.

Os indicadores de vendas (valor de vendas, quantidade de unidades e quantidade de vendas) por dia da semana na figura 12 tem um comportamento relativamente estável, mas é possível observar nuances como a venda das terças-feiras que são levemente superiores aos demais dias da semana (não considerando o final de semana), e o sábado é o melhor dia da semana em volume de vendas.

Figura 12 – Participação dos indicadores de venda por dia da semana

Fonte: Elaborado pelo autor

As quantidades de vendas e de unidades vendidas têm comportamento inverso em dois períodos distintos:

1. Nas segundas, terças, quartas e quinta-feiras o percentual de quantidade de vendas é maior que o percentual de unidades vendidas.
2. Nas sexta-feiras, o percentual de unidades vendidas é menor que o percentual de quantidade de vendas, indicando vendas de maior valor agregado no de fim-de-semana.

Nos indicadores de venda por dia do mês (figura 13) observa-se que o período próximo ao 5º dia útil tem um pico de venda e por volta do dia 15 ocorre um outro menos pronunciado. Este comportamento é explicado pelo pagamento dos salários até o 5 dia útil do mês seguinte e muitas empresas efetuam um adiantamento no dia 15 de cada mês.

Figura 13 – Participação dos indicadores de venda por dia do mês

Fonte: Elaborado pelo autor

As vendas do dia 31 não são diretamente comparáveis, pois nem todo mês possui 31 dias, e a tendência declinante é devido a menor disponibilidade financeira dos trabalhadores nos últimos dias do mês.

A sazonalidade aparece também quando se analisa os indicadores de venda por mês (figura 14) em termos percentuais, em que ocorre uma queda abrupta nos últimos meses do ano, comportamento contrário ao esperado.

Figura 14 – Participação dos indicadores de venda por mês do ano

Fonte: Elaborado pelo autor

Uma explicação é que este período corresponde ao períodos de descontinuidade administrativa (Tabela 2), hipótese sustentada pela a variação da quantidade de unidade vendidas que sofrem flutuações de maior intensidade em relação aos demais indicadores.

A existência de múltiplos fatores sazonais e a sua correspondente periodicidade são melhor tratados em modelo de previsões de séries temporais. A análise dos dados identificou os diferentes fatores sazonais e como produzem variados aspectos do comportamento das vendas. Eles são importantes para o administrador identificar padrões e tomar atitudes gerenciais, corrigindo desvios ou ampliando efeitos positivos nas vendas.

5.3 Análise temporal dos dados por departamento

Os departamentos são o primeiro grau de subdivisão na hierarquia de dados da loja. Os dados foram agrupados por ano, mês e departamento, e em seguida foram calculadas os atributos definidos na tabela 7.

5.3.1 Dados gerais

A participação das vendas por departamento mês a mês em relação ao total do faturamento do período analisado é exibido na figura 15, sendo possível identificar a tendencia declinante das vendas de perecíveis, e a perda da posição de departamento mais importante para o de não perecíveis.

O departamento de *drugstore* manteve-se estável por todo o período, mas devido a sua baixa participação não é possível extrair informações adicionais sobre o comportamento das vendas. O departamento de perecíveis possui um comportamento mais irregular, com maiores variações entre um mês e outro em comparação ao departamento de não perecíveis.

Figura 15 – Participação das vendas mensais por departamento

Fonte: Elaborado pelo autor

A figura 16, apresenta comportamento semelhante ao mesmo gráfico elaborado para análise da loja (figura 8), e apresenta 3 pontos bastante interessantes:

1. O aumento das vendas é de inteira responsabilidade do departamento de não-perecíveis;
2. Aumento expressivo de unidades vendidas de produtos perecíveis;
3. As vendas do departamento de drugstore são inexpressivas em termos percentuais de quantidade de unidades vendidas.

Figura 16 – Participação da quantidade de unidades vendidas por departamento

Fonte: Elaborado pelo autor

Espera-se que no detalhamento da análise por seção e grupo surjam as respostas para o comportamento das vendas de não-perecíveis no final do período analisado.

Na figura 17 temos uma repetição mais clara do fenômeno identificado no gráfico anterior (figura 16), onde:

1. As quantidades de unidades vendidas do departamento de perecíveis estão em franca tendência de queda;
2. Existindo uma clara tendência de aumento do hiato entre os departamentos;
3. As vendas do departamento de drugstore não são relevantes;
4. Os períodos de descontinuidade administrativa exibidos na Tabela 2, podem explicar parcialmente o comportamento destes dados

Figura 17 – Participação da quantidade de vendas por departamento

Fonte: Elaborado pelo autor

Os gráficos indicam consistência com os gráficos anteriores, e demonstram a existência de um comportamento distinto entre os departamentos para os indicadores de venda em questão.

Como alternativa para a análise dos dados foram elaborados gráficos (figura 18) que mostram o comportamento dos indicadores de valor de venda, quantidade de vendas e quantidade de unidades vendidas para cada um dos departamentos.

Esta visão corrobora os dados identificados anteriormente:

- a) Comportamento anômalo da quantidade de itens vendidos nos departamentos de perecíveis e não perecíveis;
- b) No departamento de perecíveis os indicadores de valor de venda e de quantidade de vendas tem uma tendência de convergência para uma mesma faixa de valor;
- c) No departamento de não perecíveis os indicadores de valor de venda e quantidade de vendas têm uma tendência de divergência;

Figura 18 – Visão geral - Indicadores de vendas por departamento

Fonte: Elaborado pelo autor

- d) O departamento drogaria não tem relevância nos resultados e também não apresenta desvios nos dados exibidos.

5.3.2 Comportamento dos departamentos - relevância

Uma visão alternativa explorada foi mostrar o comportamento do valor das vendas mensais por departamento e total, para detectar tendências ou padrões (Figura 19).

Figura 19 – Variação dos indicadores de vendas por departamento

	departamento	min_qtde_vendas	min_qtde_unidades	min_valor_venda	max_qtde_vendas	max_qtde_unidades	max_valor_venda	qtde_vendas_var_perc	qtde_unidades_var_perc	valor_var_perc
1	DRUGSTORE	921	1034.000	3512.88	1464	1651.015	5803.53	58.95765	59.67263	65.20718
2	NAO PERECIVEIS	14513	23504.645	49159.36	21692	102949.448	90625.46	49.46600	337.99618	84.35037
3	PERECIVEIS	11335	9830.751	57673.10	21618	29747.497	100568.98	90.71901	202.59638	74.37762

Fonte: Elaborado pelo autor

Nos dados acima chamam a atenção:

1. A grande variação nos dados referentes as quantidades de unidades vendidas (coluna qtde_unidades_var_perc);
2. O departamento de drugstore não tem relevância, nos quesitos referentes a quantidades vendidas e valores de faturamento. Porém deve ser analisado o resultado financeiro do departamento;
3. A variação do faturamento identificada não é comparável com a a inflação do período, pois estão sendo avaliadas a variação do pior caso para o melhor caso e não a variação do período.

A Figura 20 mostra o Comparativo de faturamento por departamento e total, sintetizando as observações anteriores: Irrelevância do departamento de drugstore; Tendência declinante das vendas de perecíveis; Tendência ascendente das vendas de não-perecíveis;

Figura 20 – Vendas mensais por departamento e total

Fonte: Elaborado pelo autor

5.4 Análise temporal dos dados por seção

Os dados foram agrupados por ano, mês, departamento e seção, calculados os atributos definidos na Tabela 7.

5.4.1 Dados gerais

Resumo dos dados por seção: As vendas estão concentradas em três seções que respondem por aproximadamente 50% do total de vendas, e 7 das 14 seções não atingem 5% de participação no volume de vendas (Tabela 8).

Tabela 8 – Resumo dos dados por seção

	departamento	secao	qtde_vendas	qtde_unidades	valor_venda	qtde_vendas_perc	qtde_unidades_perc	valor_perc
1	NAO PERECIVEIS	BAZAR	39845	131650.32	239278.2	1.896018	4.118113	2.789058
2	NAO PERECIVEIS	BEBIDAS	305165	514153.01	1338435.3	14.521225	16.083063	15.600976
3	NAO PERECIVEIS	BISCOITOS	109557	126687.00	260871.8	5.213251	3.962857	3.040756
4	NAO PERECIVEIS	BOMBONIERE	137373	845097.49	341122.7	6.536871	26.435236	3.976171
5	PERECIVEIS	CARNES	156819	157646.73	1836941.6	7.462206	4.931299	21.411630
6	NAO PERECIVEIS	CONFEITARIA	55845	67998.05	204695.0	2.657375	2.127026	2.385952
7	PERECIVEIS	CONGELADOS	39613	70782.71	248918.5	1.884978	2.214132	2.901426
8	DRUGSTORE	CUIDADOS PESSOAIS	67896	77383.02	274754.2	3.230820	2.420595	3.202570
9	PERECIVEIS	FRIOS	219627	142939.62	976195.4	10.450914	4.471250	11.378661
10	PERECIVEIS	HORTIFRUTI	286419	224681.18	717752.8	13.629200	7.028183	8.366220
11	NAO PERECIVEIS	LIMPEZA	74167	80050.00	276883.0	3.529224	2.504020	3.227384
12	NAO PERECIVEIS	MATINAIS	87039	102262.00	390531.0	4.141736	3.198826	4.552081
13	NAO PERECIVEIS	MERCEARIA	242433	296124.14	783039.0	11.536134	9.262969	9.127205
14	PERECIVEIS	PADARIA	279712	359404.78	689758.2	13.310048	11.242431	8.039911

Fonte: Elaborado pelo autor.

O gráfico do indicador de vendas em termos percentuais por seção (Figura 21), corrobora os aspectos anteriormente apontados na análise dos dados por departamento, destacam-se visualmente a Seção CARNES (exibida na cor verde) que apresenta uma tendência declinante nas vendas, e a BEBIDAS (exibida na cor laranja) que apresenta tendência ascendente.

Figura 21 – Participação das vendas por seção nas vendas totais do período

Fonte: Elaborado pelo autor

A variação identificada na quantidade de unidades vendidas por loja e por departamento é exibida na Figura 22, visualmente é possível destacar: A Seção BOMBONIERE (exibida na cor laranja), apresenta tendência ascendente, e a Seção MERCEARIA (exibida na cor roxa), destaca-se por apresentar um grande crescimento nos últimos meses.

Figura 22 – Participação das quantidades de unidades vendidas por seção no total de unidades vendidas no período

Fonte: Elaborado pelo autor

A análise da variação dos dados da quantidade de vendas por loja e por departamento é exibida na Figura 23, visualmente é possível destacar:

- Seção HORTIFRUTI (exibida na cor verde) apresenta um tendência declinante,
- Seção BEBIDAS (exibida na cor laranja) apresenta tendência ascendente.

Figura 23 – Participação das quantidades de vendas por seção no total de vendas do período

Fonte: Elaborado pelo autor

O quadro resumo de variação de indicadores por seção (tabela 9), mostra o comportamento dos indicadores de valor de venda, quantidade de vendas e quantidade de unidades vendidas para o departamento e seção.

Tabela 9 – Variação das vendas por seção

2	BEBIDAS	3550	6221.000	13229.77	7749	14080.000	30692.13	1177.80090	120.33017	176.39456
3	BISCOITOS	1443	1699.000	3362.60	2426	2812.000	5786.60	68.12197	65.50912	72.08708
4	BOMBONIERE	1015	1320.000	1904.15	4573	78897.000	12555.74	350.54187	5877.04545	559.38818
5	CARNES	1842	1709.656	23562.98	3545	3670.846	42990.36	92.45385	114.71255	82.44874
6	CONFEITARIA	599	687.000	1673.89	1231	1625.000	5594.79	105.50918	136.53566	234.23881
7	CONGELADOS	282	528.000	1426.57	1575	2543.000	9774.45	458.51064	381.62879	585.17143
8	CUIDADOS PESSOAIS	921	1034.000	3512.88	1464	1651.015	5803.53	58.95765	59.67263	65.20718
9	FRIOS	2805	1787.531	12199.96	4685	2963.550	22306.38	67.02317	65.79013	82.83978
10	HORTIFRUTI	3089	2118.311	6851.06	6504	5640.905	16582.81	110.55358	166.29258	142.04736
11	LIMPEZA	883	928.000	3519.73	1686	1801.000	6193.77	90.93998	94.07328	75.97287
12	MATINAIS	1088	1212.000	4844.51	2044	2453.000	9377.47	87.86765	102.39274	93.56901
13	MERCEARIA	3264	3974.678	10363.46	5353	6540.811	17414.83	64.00123	64.56203	68.04069
14	PADARIA	2889	2273.895	7187.00	6446	21315.365	16134.58	123.12219	837.39443	124.49673

Fonte: Elaborado pelo autor.

A existência de seções com alta variação é um indicador da possível ocorrência de alguns destes problemas:

- sazonalidade nas vendas,
- erro na codificação dos produtos
- falta do produto nas gôndolas.

Uma visualização alternativa do conjunto de dados em forma de Treemap (Figura 24), permite a identificação visual da importância das seções. Destacam-se as seções de carnes e frios que possuem baixa quantidade de unidades vendidas e ocupam a 1ª e 3ª posições no quesito valor de venda, respectivamente.

Figura 24 – Treemap dos indicadores de vendas

Fonte: Elaborado pelo autor

5.4.2 Principais seções

A Figura 25 permite uma visão do comportamento dos indicadores de vendas para Participação das três principais seções no período.

Figura 25 – TOP 3 - seção mais relevantes

Fonte: Elaborado pelo autor

A Seção CARNES, possui uma queda constante da participação do valor das vendas, a seção BEBIDAS mostra uma alta flutuação dos indicadores, associados a uma inversão de posições entre a quantidade de unidades e o valor de venda, e a seção FRIOS, mostra estabilidade nos indicadores com um viés de alta para o valor de vendas e um viés de baixa para o indicador de quantidade de unidades vendidas.

5.4.3 Seções sem relevância

Mesmo as seções com baixa participação nos indicadores possuem particularidades que podem auxiliar na formulação de ações para incrementar vendas ou remover produtos para melhores resultados operacionais. As três seções menos relevantes tem comportamento bastante diverso quando analisadas isoladamente. Na figura 26 é possível visualizar estes comportamentos.

Figura 26 – BOTTOM 3 - Seções menos relevantes

Fonte: Elaborado pelo autor

A seção de congelados possui picos no final de ano, e a seção de bazar possui picos de vendas nos anos de 2013 e 2014 que não se repetiram, indicando comportamento sazonal. A seção de confeitaria manteve-se estável em todo o período analisado.

5.5 Análise temporal dos dados por grupo

Após o agrupamento dos dados por ano, mês, departamento, seção e grupo, sendo calculado os atributos definidos na Tabela 7 e elaborados os gráficos e tabelas que demonstram as seguintes situações:

5.5.1 Dados gerais

Resumo dos indicadores por grupo: O grande número de grupos existentes é um indicador da complexidade de se gerenciar um empreendimento como o SuperX, com limitações humanas e materiais. A figura 27 mostra os valores dos indicadores para cada um dos grupos.

Figura 27 – Resumo dos dados por grupo de produtos

Departamento	Seção	Grupo	qtda_vendas	qtda_utilizacoes	qtda_ventos	qtda_utilizacoes_ventos	qtda_utilizacoes_ventos	qtda_utilizacoes_ventos
1	PEREVEIS	CARNES	8661	8649.254	1424341.30	4.26402069	2.762364732	14.46051611
2	PEREVEIS	PADARIA	246072	305346.977	604305.51	11.709294745	9.5514652033	7.043636375
3	NÃO PEREVEIS	BEBIDAS	REFRIGERANTE	131103	102183.000	562712.85	6.238514211	4.760388111
4	PEREVEIS	FRIGOS	EMBUTIDOS	100143	44074.412	200706.40	5.002511976	3.770420040
5	PEREVEIS	CARNES	AVES	53008	60889.457	488355.03	2.546169183	1.90466464547
6	NÃO PEREVEIS	BEBIDAS	CERVICAS	47353	186789.000	423452.74	2.252384543	3.8432023346
7	PEREVEIS	HORTIFRUTI	LEGUMES	127145	79144.008	26881.83	6.314846605	2.146147140
8	NÃO PEREVEIS	BOMBONIERE	DOCES	97030	794975.486	206626.18	4.620164695	24.8548742682
9	PEREVEIS	HORTIFRUTI	FRUTAS	70984	66624.686	203487.59	3.377811105	2.080386388
10	PEREVEIS	FRIGOS	QUEIJO	40756	12865.381	198612.42	2.177243862	3.3686705177
11	NÃO PEREVEIS	MATINAS	LEITE	44164	55458.000	182678.38	2.101530514	1.734764119
12	NÃO PEREVEIS	BEBIDAS	REFRESCO	83652	136481.000	178893.37	4.408829841	4.2892203084
13	PEREVEIS	CONGELADOS	SOPAS	23942	28609.000	168860.19	1.182676444	0.928190601
14	NÃO PEREVEIS	BAZAR	UTILIDADES	23027	110975.700	188158.15	1.067359233	3.4620126730
15	NÃO PEREVEIS	BISCOTOS	DOCE	65288	76459.000	138893.68	3.108718502	2.3919506005
16	NÃO PEREVEIS	MERCERIA	SALGADOS	46258	53107.788	137814.47	2.200120286	1.8615476568
17	PEREVEIS	HORTIFRUTI	VERDURAS	56610	60385.423	120178.17	2.653777332	1.888819215
18	DRUGSTORE	CUIDADOS PESSOAS	DESCARTAVES	22219	23072.000	115207.10	1.057387379	1.342806014
19	NÃO PEREVEIS	LIMPEZA	SAPONIFICOS	23627	21538.000	110387.88	1.419121280	0.984875989
20	PEREVEIS	FRIGOS	DOUR	34969	42076.180	111987.96	1.644879754	1.547897602
21	NÃO PEREVEIS	MATINAS	CAFE	19088	19520.000	110851.44	0.908204101	0.6105991412
22	PEREVEIS	HORTIFRUTI	OVOS	35702	27228.000	108818.54	1.225202348	0.8517107584
23	NÃO PEREVEIS	BOMBONIERE	CHOCOLATES	25302	31421.000	97218.86	1.205274551	0.9831544460
24	NÃO PEREVEIS	CONFETARIA	SORREMESAS	26749	34587.000	96150.80	1.272846667	1.082181603
25	PEREVEIS	CARNES	BURR	13650	7824.021	95541.07	1.846520205	0.244240134
26	NÃO PEREVEIS	MERCERIA	MAGARIM	36256	84452.000	95154.03	1.725235954	0.7186238730
27	PEREVEIS	PADARIA	LANCHES	33640	54037.802	85462.68	1.600733744	1.6906505471
28	NÃO PEREVEIS	MERCERIA	FARMACIOS	24850	28802.389	84268.65	1.187214555	0.8348701602
29	NÃO PEREVEIS	CONFETARIA	BOLOS	13318	26418.000	81632.01	1.108630894	0.826420678
30	NÃO PEREVEIS	MERCERIA	TEMPEROS	27564	30469.000	71208.13	1.311828305	0.952778468
31	NÃO PEREVEIS	MERCERIA	EMBALADOS	19332	23323.000	60886.22	0.922760509	0.729598194
32	NÃO PEREVEIS	BEBIDAS	DESTILADOS	5842	10144.000	5846.36	0.438779016	0.3176414335
33	NÃO PEREVEIS	MERCERIA	ARROZ	8198	9095.000	59678.67	0.339104052	0.284487903
34	NÃO PEREVEIS	MERCERIA	MOLCOS	26396	32484.000	58889.14	1.20589188	0.916122045
35	PEREVEIS	FRIGOS	MANGARIM	16015	16918.000	58851.79	0.78585408	0.521090487
36	NÃO PEREVEIS	MATINAS	BEBIDAS	14681	17260.000	53313.12	0.688892617	0.539647735
37	PEREVEIS	BAZAR	DESCARTAVES	13246	14648.000	49691.39	0.632038683	0.496383196
38	NÃO PEREVEIS	MERCERIA	FEIJAO	7616	8485.075	46857.84	0.363466678	0.278303075
39	NÃO PEREVEIS	MERCERIA	ACUCAR	14206	19011.000	48328.11	0.680227756	0.594872581
40	NÃO PEREVEIS	MERCERIA	OLEOS AZEITES	10643	11720.000	46949.03	0.50643365	0.366097303
41	NÃO PEREVEIS	BISCOTOS	CRISUM CRACKER	16911	16020.000	45075.04	0.846701052	0.6117271106
42	DRUGSTORE	CUIDADOS PESSOAS	PERFUMARIA	14750	19181.000	33342.96	0.701876270	0.599649661
43	DRUGSTORE	CUIDADOS PESSOAS	ABSORVENTE	8352	10188.000	33318.58	0.440248684	0.318938005
44	NÃO PEREVEIS	LIMPEZA	ALVEJANTES / AGUA SANITARIA	8193	8223.000	31970.63	0.452147288	0.291690678
45	PEREVEIS	FRIGOS	BEBIDAS	12503	20694.000	30798.07	0.59465153	0.653578144
46	NÃO PEREVEIS	LIMPEZA	AMACIANTES	5918	6026.000	29752.83	0.28151865	0.184847403
47	NÃO PEREVEIS	BEBIDAS	ESPUMANTE	4387	4803.000	29732.45	0.276754657	0.153264513
48	NÃO PEREVEIS	LIMPEZA	LIMPADORES	8913	7113.000	28924.03	0.32893843	0.222499744
49	NÃO PEREVEIS	BEBIDAS	AGUA	13014	15028.000	28778.99	0.642061418	0.498238884
50	NÃO PEREVEIS	BOMBONIERE	DOCES	15849	12762.000	28787.81	0.49816488	0.4001426340
51	DRUGSTORE	CUIDADOS PESSOAS	HIGIENE ORAL	8644	9652.000	27424.68	0.432537119	0.310468681
52	NÃO PEREVEIS	BISCOTOS	WAFER	10024	12508.000	26586.78	0.518816703	0.3840741794
53	PEREVEIS	CONGELADOS	HAMBURGUES	8525	25140.000	25958.67	0.319401028	0.1785396339
54	DRUGSTORE	CUIDADOS PESSOAS	TRATAMENTO CAPELO	3398	3364.000	25515.16	0.158494775	0.105226536
55	NÃO PEREVEIS	BEBIDAS	VINHO	3184	3365.000	25333.88	0.158101005	0.105226534
56	NÃO PEREVEIS	LIMPEZA	UTILIDADES	11716	14816.000	22297.47	0.55750380	0.4534747503
57	DRUGSTORE	CUIDADOS PESSOAS	BABERIA	9558	807.000	22031.44	0.275791942	0.2523291748
58	NÃO PEREVEIS	MATINAS	COMIDAS	4128	4501.000	22196.92	0.196420186	0.140882445
59	NÃO PEREVEIS	BAZAR	RACAO	3572	5069.615	21429.71	0.169973019	0.150206879
60	PEREVEIS	CONGELADOS	APERTIVOS	4712	10615.515	21247.47	0.254719728	0.3388138417
61	PEREVEIS	FRIGOS	REQUEIJAO	3086	3169.668	17513.38	0.146890594	0.088362041
62	NÃO PEREVEIS	BEBIDAS	APERTIVOS	1810	2025.000	17195.56	0.098128546	0.0633434048
63	NÃO PEREVEIS	MERCERIA	MANGARIM	4406	5104.000	16277.27	0.222007013	0.159666607
64	PEREVEIS	CONGELADOS	MASSA	2132	2038.500	14920.41	0.10140862	0.0794065410
65	NÃO PEREVEIS	MATINAS	CERIAS	2344	3340.000	14787.81	0.144848228	0.1044737170
66	NÃO PEREVEIS	MERCERIA	CARPAO	1958	2733.000	14769.97	0.08511101	0.064061339
67	NÃO PEREVEIS	BISCOTOS	SALGADO	4520	4991.000	14686.53	0.215084440	0.1561219423
68	NÃO PEREVEIS	CONFETARIA	MISTURA PARA BOLO	3210	4681.000	13951.22	0.152747110	0.1276686110
69	NÃO PEREVEIS	LIMPEZA	DETERGENTES	1616	1701.000	12702.02	0.078607188	0.0532584600
70	NÃO PEREVEIS	BISCOTOS	ROSQUINHA	3274	3400.000	12680.82	0.158792740	0.1063543386
71	NÃO PEREVEIS	BEBIDAS	VAGLIMARE	1792	2198.010	11583.15	0.089272019	0.067186868
72	NÃO PEREVEIS	BISCOTOS	AGUA E SAL	4633	5289.000	11575.10	0.222460206	0.1697360900
73	NÃO PEREVEIS	BISCOTOS	MAIZENA	4007	4320.000	11370.44	0.190674241	0.1351329568
74	NÃO PEREVEIS	LIMPEZA	DESMONTANTES	3219	3268.000	11244.63	0.153179374	0.1021627457
75	NÃO PEREVEIS	MERCERIA	CONSERVAS	2888	3040.010	11118.07	0.12700643	0.099396217
76	NÃO PEREVEIS	CUIDADOS PESSOAS	DESCOORDANTES	1148	1192.000	10746.08	0.054607387	0.0372866869
77	NÃO PEREVEIS	MERCERIA	SAL	4991	5160.000	9722.39	0.237489896	0.1620339934
78	PEREVEIS	CONGELADOS	VEGETAIS	1511	1783.000	9391.54	0.071900881	0.0557374769
79	NÃO PEREVEIS	MERCERIA	SOPAS/COMIDA INST	2038	2611.000	9137.90	0.068977887	0.0616738913
80	NÃO PEREVEIS	LIMPEZA	EQUIPAMENTOS	3066	2191.000	8892.13	0.098310263	0.068339999
81	PEREVEIS	CONGELADOS	COMIDA PRONTA	704	847.000	7970.13	0.02584812	0.028484178
82	NÃO PEREVEIS	CONFETARIA	SABONAL	483	659.000	7872.94	0.022864574	0.0174850678
83	PEREVEIS	FRIGOS	SORREMESAS	1778	1998.000	6949.81	0.084510681	0.0624988260
84	NÃO PEREVEIS	MERCERIA	CONCOMIDOS	2382	3501.000	6618.88	0.160391901	0.111500003
85	NÃO PEREVEIS	MERCERIA	ALCOOL	1281	1347.000	6343.41	0.060969170	0.0421350044
86	DRUGSTORE	CUIDADOS PESSOAS	PEMBO	2138	2268.000	5303.97	0.101736271	0.069697717
87	NÃO PEREVEIS	CONFETARIA	FERNITO	2084	2342.000	4907.06	0.091667030	0.072593587
88	NÃO PEREVEIS	BOMBONIERE	SALGADOS	1890	3127.000	4602.96	0.098930332	0.0871447482
89	NÃO PEREVEIS	MERCERIA	LAMPADOR CIBRILHO	839	855.000	4366.09	0.030626274	0.026146931
90	NÃO PEREVEIS	LIMPEZA	DESCOORDANTES	1867	1847.000	4066.54	0.086000192	0.051593025
91	NÃO PEREVEIS	MERCERIA	PIZZAS	2127	2742.000	3644.24	0.017125327	0.087178622
92	NÃO PEREVEIS	MATINAS	ADOCANTES	962	1024.000	3524.05	0.04779608	0.0320143024
93	NÃO PEREVEIS	MATINAS	CREME	974	1165.000	3370.31	0.045347626	0.036104419
94	NÃO PEREVEIS	MERCERIA	MISCELANEA	707	862.000	3286.32	0.032642676	0.0207176412
95	PEREVEIS	HORTIFRUTI	ALHO	378	399.000	1406.64	0.01878066	0.0124839968
96	DRUGSTORE	CUIDADOS PESSOAS	SEM ESTAR BEBE	189	199.000	1269.87	0.008930333	0.008248880
97	NÃO PEREVEIS	BEBIDAS	ESPUMANTE	128	148.000	1106.48	0.00606068	0.0045662020
98	PEREVEIS	FRIGOS	SORREMESAS	368	375.000	1055.95	0.014666138	0.011203601
99	NÃO PEREVEIS	BOMBONIERE	BALAS	382	430.000	834.09	0.018177406	0.013450863
100	NÃO PEREVEIS	LIMPEZA	TECIDOS SACOS PANOS	194	180.000	785.65	0.008331467	0.0071632788
101	NÃO PEREVEIS	CONGELADOS	PEIXES	27	30.000	533.17	0.001284790	0.000384208
102	NÃO PEREVEIS	MERCERIA	FRUTAS	32	38.000	133.16	0.001522715	0.001186964

Fonte: Elaborado pelo autor

A Figura 28 mostra o comportamento dos indicadores estudados, identifica a coerência com os gráficos anteriores e detalha como o conjunto dos grupos impactam uma seção.

Figura 28 – Visão geral dos grupos

Fonte: Elaborado pelo autor

A análise dos gráficos permite identificar que os maiores valores para os indicadores estudados estão concentrados em algumas poucas seções.

A Figura 29 apresenta um resumo da variação dos indicadores de venda por grupo de produtos, o comportamento dos valores extremos (valor mínimo, valor máximo e a respectiva variação percentual) para estes indicadores (valor de venda, quantidade de vendas e quantidade de unidades vendidas).

Figura 29 – Variação das vendas por GRUPO

	grupo	min_qtde_vendas	min_qtde_unidades	min_valor_venda	max_qtde_vendas	max_qtde_unidades	max_valor_venda	qtde_vendas_var_perc	qtde_unidades_var_perc	valor_var_perc
1	ABSORVENTE	123	130.000	422.32	206	233.000	741.84	67.47967	79.23077	75.65827
2	ACUCAR	179	213.000	451.47	406	671.000	1737.97	126.81564	215.02347	284.95803
3	ADOCANTES	8	8.000	27.99	27	30.000	148.15	237.50000	275.00000	429.29618
4	AGUA	98	110.000	228.33	465	585.000	1093.52	374.48980	431.81818	378.52086
5	AGUA E SAL	29	31.000	88.99	172	198.000	400.48	493.10345	538.70968	345.02723
6	ALCOOL	7	7.000	35.23	37	43.000	196.16	428.57143	514.28571	456.79818
7	ALHO	2	2.000	4.00	51	56.000	146.07	2450.00000	2850.00000	3075.43478
8	ALVEJANTES / AGUA SANITARIA	106	106.000	359.30	214	220.000	802.86	101.88879	103.70370	123.45116
9	AMACIANTE	65	66.000	337.73	149	151.000	718.70	129.23077	128.76788	112.80313
10	APERITIVOS	18	19.000	153.39	186	381.000	1512.57	933.33333	1905.26316	886.09427
11	ARROZ	116	126.000	760.62	180	189.000	1390.85	55.17241	50.00000	82.85741
12	AVES	710	799.630	6980.22	1203	1511.262	11696.30	69.43662	88.99516	67.56349
13	BALAS	1	1.000	1.50	40	43.000	57.43	3900.00000	4200.00000	3728.66667
14	BARBEARIA	52	71.000	189.58	181	250.000	630.78	248.07692	252.11268	232.72497
15	BEBIDAS	281	378.000	854.56	629	970.000	1962.82	123.84342	196.61376	132.02818
16	BEM ESTAR BEBE	1	1.000	2.50	28	28.000	181.15	2700.00000	2700.00000	7146.00000
17	BOLOS	226	247.000	675.85	601	713.000	1934.47	165.92920	188.66367	186.22771
18	BOVINA	981	842.496	15069.06	2121	2519.939	30663.26	116.20795	199.10397	103.48489
19	CAFE	236	236.000	1167.87	480	496.000	2833.52	103.38983	107.53138	142.62289
20	CARVAO	17	24.000	119.76	66	113.000	563.87	288.23529	370.83333	370.83333
21	CEREAIS	21	27.000	117.63	93	103.000	482.27	342.85714	281.48148	309.98895
22	CERVEJAS	513	1464.000	3016.57	1231	7454.000	15342.05	139.96101	409.15301	408.59254
23	CHOCOLATES	223	259.000	859.44	727	924.000	2788.22	226.00997	256.75676	225.58643
24	COMIDA PRONTA	2	2.000	15.96	37	42.000	380.82	1750.00000	2050.00000	2278.99762
25	COMIDAS	42	43.000	227.36	104	130.000	599.05	147.61905	202.32559	159.06251
26	CONDIMENTOS	26	26.000	48.36	103	109.000	213.52	296.15385	319.23077	341.52192
27	CONSERVAS	24	26.000	87.60	124	146.000	449.81	416.66667	461.53846	413.48174
28	CREAM CRACKER	163	194.000	395.13	459	532.000	1240.13	181.59509	174.22680	213.85367
29	CREME	1	1.000	2.19	63	77.000	186.10	6200.00000	7600.00000	8397.71689
30	DESCARTAVEIS	478	526.000	1945.86	813	887.000	3816.92	70.08368	68.63118	96.15594
31	DESINFETANTES	29	30.000	115.32	115	115.000	344.85	296.55172	283.33333	199.03746
32	DESCORDOANTES	2	2.000	11.50	48	50.000	352.17	2300.00000	2400.00000	2962.34783
33	DESCORDOADORES	8	8.000	16.23	52	63.000	187.99	550.00000	687.50000	935.95853
34	DESTILADOS	95	102.000	577.82	243	299.000	1746.46	155.78947	193.13725	202.24984
35	DOCE	793	957.000	1932.49	1873	2182.000	3728.13	136.19168	128.00418	92.91846
36	DOCES	606	842.000	797.30	3564	78179.000	9183.76	488.11881	9184.91686	1051.85752
37	EMBUTIDOS	1411	952.601	6556.75	2397	1044.468	12903.95	69.87952	76.25147	96.80406
38	ENLATADOS	175	201.000	552.09	480	602.000	1461.63	174.28571	199.50249	164.74488
39	EQUIPAMENTOS	12	12.000	57.66	72	77.000	327.83	500.00000	541.66667	468.55706
40	ESPUMANTES	1	1.000	6.50	19	24.000	159.36	1800.00000	2300.00000	2351.69231
41	ESTIMULANTES	25	26.000	160.56	155	182.000	1084.34	560.00000	600.00000	581.57668
42	FARINACEOS	295	322.648	883.48	588	787.434	2123.71	99.32203	144.05358	140.36009
43	FELJAO	91	100.000	427.28	200	231.000	1920.75	119.78022	131.00000	349.52558
44	FERMENTO	14	14.000	30.78	63	73.000	164.13	350.00000	421.42857	433.23587
45	FRUTAS	749	528.188	2003.63	1614	1589.517	4746.68	115.48732	200.93773	136.90402
46	HAMBURGUER	37	134.000	205.74	182	696.000	681.32	391.89189	419.40299	231.15583
47	HIGIENE ORAL	108	114.000	310.41	208	229.000	636.05	92.59259	100.87719	104.90641
48	INSETICIDAS	1	1.000	12.89	105	112.000	916.28	10400.00000	11100.00000	7008.46617
49	ISOLANT	272	332.000	924.75	817	1134.000	2694.87	200.36765	240.54054	186.16275
50	LANCHES	169	277.550	498.93	1019	3148.596	2815.42	502.95858	1033.70420	478.19810
51	LEGUMES	1650	765.815	2795.96	3000	1612.475	7905.80	81.81818	110.55673	172.02821
52	LEITE	586	693.000	2312.80	1057	1393.000	4688.77	80.37543	101.01010	102.73132
53	LIMPADOR C/BRILHO	3	3.000	12.77	24	25.000	134.77	700.00000	733.33333	955.36413
54	LIMPADORES	81	83.000	368.09	185	188.000	674.42	128.39506	126.50602	83.22149
55	MACARRAO	524	781.000	1206.09	818	1229.000	2094.59	56.10687	57.36236	73.66780
56	MAIONESES	37	39.000	154.50	141	175.000	559.78	281.08108	348.71795	262.31715
57	MAIZENA	36	39.000	99.58	118	129.000	343.48	227.77778	230.76923	244.52870
58	MARGARINA	207	209.000	671.30	392	394.000	1498.37	88.37198	101.14518	121.71658
59	MASSA	6	6.000	80.44	77	87.000	555.03	1183.33333	1350.00000	589.99254
60	MASSAS	2	3.000	7.17	39	70.000	207.03	1850.00000	2233.33333	2787.44770
61	MISTURA PARA BOLO	30	38.000	101.52	83	108.000	362.62	176.66667	184.21053	257.19070
62	MOLHOS	307	367.000	693.65	713	898.000	1474.59	132.24756	144.68665	112.58416
63	OLEOS AZEITES	143	154.000	607.56	242	270.000	1110.55	69.23077	75.32468	82.78853
64	OVOS	170	222.000	828.43	627	894.000	2585.37	268.82353	302.70270	212.08068
65	PAES	2720	1762.400	6402.38	5629	20970.000	14271.07	106.54853	1072.83250	122.90258
66	PEMAO	7	7.000	26.93	71	914.28571	181.40	914.28571	928.57143	796.69892
67	PEIXES	1	2.000	23.98	5	6.000	94.85	400.00000	200.00000	179.15479
68	PERFUMARIA	161	238.000	419.67	347	443.000	751.76	115.52795	86.13445	79.13122
69	PIPOCAS	3	3.000	2.97	90	118.000	158.91	2900.00000	3833.33333	5250.50505
70	QUEIJOES	572	147.972	2400.85	1058	287.683	4509.18	84.96503	94.41719	87.81598
71	RACAO	7	10.000	33.71	126	184.015	842.86	1700.00000	1740.15000	2400.23631
72	REFRESCO	1070	1698.000	2357.56	2480	3829.000	4694.72	131.77570	125.50059	99.13470
73	REFRIGERANTE	1365	1472.000	6151.36	3353	4080.000	15963.31	145.64103	177.17391	159.50863
74	REQUEIJAO	24	24.000	58.18	110	115.000	455.15	400.00000	422.72727	682.31351
75	ROSQUINHA	22	22.000	58.18	110	115.000	455.15	400.00000	422.72727	682.31351
76	SAL	47	49.000	83.31	117	122.000	241.04	148.93617	148.97959	189.32901
77	SALGADINHOS	580	675.000	1215.38	1170	1280.000	3286.86	101.72414	89.62963	170.43888
78	SALGADO	29	30.000	102.08	172	189.000	544.12	493.10345	530.00000	433.03292
79	SALGADOS	2	2.000	13.18	106	246.000	264.35	5200.00000	12200.00000	1905.69044
80	SAPONACEOS	358	370.000	1513.40	709	752.000	2526.46	99.15730	103.24324	66.93934
81	SAZONAL	1	1.000	11.99	57	71.000	1468.15	5600.00000	7000.00000	12144.78732
82	SOBREMESA	4	5.000	24.45	58	71.000	205.96	1350.00000	1320.00000	741.14919
83	SOBREMESAS	299	374.000	869.90	707	998.000	2871.23	138.45485	168.44492	230.44424
84	SOPAS/COMIDA INST.	6	7.000	29.03	87	117.000	418.65	1350.00000	1571.42857	1342.12883
85	SORVETES	78	100.000	636.22	1164	1511.000	7630.95	1392.30769	1411.00000	1099.42001
86	SUINA	96	51.243	675.44	350	245.352	2524.55	264.58333	378.80101	273.76377
87	TECIDOS SACOS PANOS	1	1.000	1.50	19	25.000	59.75	1800.00000	2400.00000	3883.33333
88	TEMPEROS	341	370.000	877.46	645	729.000	1626.99	89.14956	97.02703	85.42042
89	TRATAMENTO CABELO	17	17.000	161.73	104	105.000	799.93	511.76471	517.64706	394.06830
90	UTILIDADES	310	463.810	1599.52	7737	8816.321	9696.46	460.32258	1800.84755	506.21061
91	VASILHAME	1	1.000	7.99	626	750.000	3750.00	6260.00000	7490.00000	4653.86708
92	VEGETAIS	1	1.000	4.29	76	84.000	431.11	7500.00000	8300.00000	9720.27335
93	VERDURAS	516	527.025	1225.73	1517	1608.709	3567.20	193.99225	205.24339	191.02657
94	VINHO	20	22.000	132.70	99	110.000	846.61	395.00000	400.00000	537.98794
95	WAFER	72	78.000	196.52	395	513.000	1085.69	448.61111	557.69231	452.45777

Fonte: Elaborado pelo autor

O Treemap (Figura 30) também foi utilizada para a identificação de características do

conjunto de dados quando agregados por departamento, seção e grupo:

Figura 30 – Treemap dos indicadores de vendas por grupo

Fonte: Elaborado pelo autor

A inspeção visual dos gráficos de mapa de calor mostra que:

1. Há uma quantidade desproporcional do grupo doces quando em comparação aos demais;
2. O grupo legumes embora seja o 2 em quantidade de venda é apenas o 10 em quantidade de unidades vendidas;
3. A confirmação de que o grupo carne bovina é o mais importante para o resultado das vendas, seguidos pelos grupos de pães, refrigerantes e embutidos.

5.5.2 Principais grupos

A figura 31 se refere ao comportamento dos indicadores de vendas para Participação dos três principais grupos de produtos no período. Observa-se que:

1. a queda nas vendas de carnes bovinas, que impacta na seção de carnes e no departamento de perecíveis.
2. Há um aumento nas unidades vendidas de pães sem a devida contrapartida no valor das vendas, fato que deve ser investigado pois aparenta ser erro nos dados de origem e que não foi detectado no processo de engenharia de dados.
3. Ocorreu uma diminuição na quantidade de unidades vendidas de refrigerantes, sem a correspondente diminuição do faturamento, indicando um processo de reposição de margens neste produto.

Figura 31 – TOP 3 - Grupos mais relevantes

Fonte: Elaborado pelo autor

5.5.3 Grupos sem relevância

A participação dos três grupos menos relevantes no período fornecem uma visão do comportamento dos indicadores de vendas nesta situação, exibido na Figura 32.

Figura 32 – BOTTOM 3 - Grupos menos relevantes

Fonte: Elaborado pelo autor

Os grupos acima não possuem variação visualmente significativas nos indicadores, sendo possível visualizar pequenas oscilações para o grupo sazonal durante o período de gestão do proprietário nº. Devido a baixa representatividade, os produtos pertencentes a estes grupos poderiam ser excluídos das prateleiras sem prejuízo para o SuperX.

5.6 Análise temporal dos dados por produto

Os produtos são os itens ofertados pelo varejo, e as principais operações de um estabelecimento varejista são realizadas por produto, sejam elas a compra, o armazenamento, a exposição ou a venda. Atividades gerenciais como previsão de demanda, controle dos estoques ou prevenção de perdas, entre outras atividades também são realizados por produto.

O SuperX possui 9.106 diferentes produtos, e embora classificados em departamento, seção e grupos são eles individualmente que dão vida ao estabelecimento. Problemas apontados nas análises dos dados agrupados podem ser melhor compreendidos quando estudadas por produto.

5.6.1 Visão geral dos produtos

Identificar a quantidade de produtos vendidos em um período e comparar com as quantidades comercializadas em outros períodos, na busca de padrões que auxiliem na gestão do empreendimento (Figura 33):

Figura 33 – Quantidade de produtos comercializados no mês

Fonte: Elaborado pelo autor

Os períodos em que ocorrem uma maior queda na quantidade de diferentes itens vendidos (quantidade de diferentes produtos) durante o mês, é condizente com os períodos de descontinuidade administrativa (Tabela 2), quando ocorreram problemas de abastecimento no SuperX.

Na Figura 34 temos a visão inversa, que é identificar a quantidade de produtos não comercializados mês a mês, fornecendo uma perspectiva complementar à anterior, onde os picos (ou vales, no caso do gráfico anterior) representam períodos onde possivelmente ocorreram falta de mercadorias nas lojas.

Figura 34 – Quantidade de produtos não comercializados no mês

Fonte: Elaborado pelo autor

5.6.2 Principais produtos comercializados

A figura 35 traz os trinta principais produtos selecionados pelo critério de valor de vendas no período pesquisado:

Figura 35 – Lista dos 30 principais produtos

	codigoproduto	descricao completa	qtde_vendas	qtde_unidades	valor_venda
1	7852	PAO FRANCES UN	179851	209250.828	374859.22
2	19862	REFR COCA COLA RETORNAVEL 2L	40099	52437.000	196611.67
3	4118	MUSSARELA KG	37695	6798.646	164361.33
4	1456	CARNE BOVINA MIOLO DE ACEM KG	9292	10829.735	149629.64
5	7869	FRANGO FILE KG	12872	11082.998	124164.93
6	644	CERV SKOL LT 350ML	12897	44386.000	119896.98
7	3687	CARNE BOVINA COXAO MOLE KG	7307	6273.028	119780.81
8	3657	CARNE BOVINA CONTRA FILE KG	5716	5099.376	118633.75
9	3587	CARNE BOVINA MOIDA DE SEGUNDA KG	11884	9578.691	99331.77
10	69	DOCE UN	34158	93045.000	95634.43
11	3647	CARNE BOVINA ACEM KG	7427	7488.739	92872.44
12	1287	CARNE BOVINA PICADAO KG	7837	7693.496	87935.19
13	6	DIVERSOS.	9344	85325.860	85325.86
14	9235	CARNE BOVINA COSTELA GAUCHA KG	5085	8107.927	84635.76
15	9113	FRANGO A PASSARINHO KG	9057	10366.908	72447.11
16	3693	LINGUICA TOSCANA AURORA KG	6609	5290.023	69813.34
17	5939	FRANGO PEITO KG	7524	8371.198	65719.24
18	52	BALA/CHICLETE/DOCE/PIRULITO 0,10	33263	654125.000	65412.50
19	981	CARNE BOVINA COXAO DURO KG	4104	3781.871	65319.25
20	14518	CERV ITAIPAVA LT 350ML	7540	24614.000	64328.32
21	8799	FRANGO COXA E SOBRECORA KG	6450	9343.152	62021.27
22	3732	BATATA LAVADA KG	22516	21223.266	61626.13
23	32	TOMATE PIZZADORO KG	26354	14981.084	59872.11
24	3709	CARNE BOVINA PATINHO KG	3946	3230.461	55947.22
25	3923	PAO DE FORMA PANCO 500G	10440	10884.000	52915.98
26	6211	FRANGO INTEIRO MARINGA KG	3564	9332.339	52509.53
27	10	COENTRO UN	25275	26564.000	50591.50
28	2418	REFR COCA COLA PET 2L	7450	7928.000	50291.29
29	5646	ALHO ROXO KG	16247	2133.864	47365.25
30	3584	CARNE BOVINA ALCATRA KG	2290	2134.812	47045.38

Fonte: Elaborado pelo autor

Pelos valores envolvidos observa-se a necessidade crítica de precificação e estoque destes produtos, pois representam um valor percentual muito importante do faturamento, e qualquer falha tem impactos graves nos resultados.

5.6.3 Produtos sem relevância no resultado

A figura 36 traz os trinta produtos que tiveram os menores valores de vendas no período pesquisado. São produtos que foram comercializados uma única vez no estabelecimento, ou então falhas de cadastro, não possuindo nenhuma relevância para os negócios.

Figura 36 – Lista dos 30 produtos menos relevantes

	codigoproduto	descricaoCompleta	qtde_vendas	qtde_unidades	valor_venda
9077	11554	TAQUARAL PINCA	2	2.000	1.98
9078	17521	IOG DANONE ACTIVIA CEREJA 100G	2	2.000	1.98
9079	13929	ALGODAO CLEAN SOFT BABY 50G	1	1.000	1.89
9080	14388	IOG ACTIVIA AVEIA 200G	1	1.000	1.89
9081	24339	FEIJÃO FRADINHO MAXIMO TP1 500G	1	1.000	1.89
9082	19917	MORANGUETE BEL	4	7.000	1.75
9083	24199	GELATINA NEILAR 30G MORANGO	1	2.000	1.58
9084	24201	GELATINA NEILAR 30G BLUEBERRY	1	2.000	1.58
9085	10702	PRATOS DESC COPOSUL ROSA 150MM 10UN	1	1.000	1.50
9086	18526	PAO DOCE CARACOL UN	1	1.000	1.50
9087	18665	PALITO MANICUREUN	5	5.000	1.50
9088	24169	MINI CLICLETES GLOBY	1	1.000	1.50
9089	13973	SABONETE LUX LUXO SEDUZA-ME 90G	1	1.000	1.49
9090	15480	MAC ORIENTAL BIFUM 70G	1	1.000	1.49
9091	12305	TAQUARAL BUCHA Nº 10	2	2.000	1.40
9092	11878	SABONETE NIVEA HIDRATANTE BAMBU 90G	1	1.000	1.39
9093	19610	TAQUARAL BUCHA REDUCAO RR 3/4X1/2	1	1.000	1.35
9094	19634	TAQUARAL TAMPA PLAST P/ PIA PQ 2UN	1	1.000	1.00
9095	20335	BALAS DORI GOMETS FRUTAS 32G	1	1.000	1.00
9096	10457	PILHA RAYOVAC AMARELINHAS AA 1UN	1	1.000	0.99
9097	14404	MAC PANCO LAMEN CREMOSO CR DE MILHO 85G	1	1.000	0.99
9098	18669	LIXA DE U NHA	3	3.000	0.90
9099	22150	TAQUARAL BRINQUEDO	1	1.000	0.80
9100	3480	ERVILHA JUSSARA 200G	1	1.000	0.80
9101	17697	CREME DUCREM CHOCOLATE/AVELA 10G	1	2.000	0.50
9102	1207	PAOZINHO DOCE UN	1	1.000	0.30
9103	182	FEIJAO FRADINHO CALDO NOBRE KG	3	0.035	0.16
9104	3752	RACAO WHISKAS KG	1	0.005	0.05
9105	77	FEIJAO FRADAO KG	1	0.005	0.03
9106	140	TOMATE ITALIANO	1	0.005	0.01

Fonte: Elaborado pelo autor

5.7 Agrupamento de produtos

Ficou identificado que 85% das vendas são devido a apenas 1.071 produtos do total de 9.106 existentes nos dados. Para diminuir a quantidade de produtos a serem considerados foi utilizado a seguinte estratégia de agrupamento:

- a) Os produtos que representam até a medida de corte estabelecida como sendo de interesse principal para análise (no caso representam 85% das vendas), mantém a sua classificação e denominação original;
- b) Os demais produtos são classificados em ordem decrescente da seguinte forma: obtém-se a variação percentual entre os dois últimos produtos, o produto seguinte mantém o número de ordem original porém sua descrição é alterada para "produto de origem 9999", e os próximos produtos são igualmente alterados, enquanto a somatória de suas variações percentuais forem inferiores a variação percentual entre o último e o penúltimo produto.

5.8 Observações

Os dados fornecidos pelo SuperX possuem muitas observações e alta dimensionalidade (grande número de variáveis). Foi necessário realizar vários recortes para se obter uma visão multifacetada destes dados, e dentro de um processo de KDD encontrar questões a serem exploradas e esclarecidas.

Para a caracterização dos dados utilizando estatística descritiva foram utilizados os programas desenvolvidos para a atividade de engenharia de dados, sendo codificados os seguintes novos programas:

Tabela 10 – Lista de artefatos - Código fonte

Programa	Descrição resumida
P020_DefineAtributos_V1	Identifica a classe de dados dos atributos.
P021_análiseEstatistica_V3	Produz análises estatísticas que descrevem as principais características das variáveis disponíveis nos dados analisados.

Fonte: Elaborado pelo autor

APRENDIZADO DE MÁQUINA

No varejo supermercadista, como em inúmeros outros ramos de negócio, uma das maiores preocupações do administrador é conseguir visualizar com antecedência eventos que interferem no resultado do negócio, para tomadas de decisão em tempo de correção e ajuste para o bom andamento do negócio. As principais preocupações estão voltadas para o aspecto financeiro, sendo a previsão do valor do faturamento a central. Outro grande ponto é o abastecimento da loja com os produtos adequados e na quantidade adequada, pois se a falta de produtos provoca perda de faturamento, o excesso é responsável por perdas financeiras, relativas ao custo de manter em estoque um produto que não tem consumidores.

De forma complementar aos pontos citados acima, existe a preocupação de se conhecer o comportamento futuro dos preços dos produtos, ter a visão destes preços futuros permite ao administrador estabelecer políticas de substituição de produtos nas gondolas, formação de estoques com fins especulativos ou a decisão de queimar estoques em operações de marketing com a missão de diminuir futuros prejuízos em função de uma esperada queda de preços.

6.1 Conceitos básicos

As técnicas de Aprendizado de Máquina (Machine Learning) empregam um princípio de inferência denominado indução, no qual obtém-se conclusões genéricas a partir de um conjunto particular de exemplos. O aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não-supervisionado.([FACELI; OUTROS, 2017](#))

O aprendizado supervisionado ocorre quando existe a simulação de um "supervisor externo" que conhece a saída desejada para cada exemplo, o objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente.([FACELI; OUTROS, 2017](#))

Um aspecto importante para o Aprendizado de Máquina é a capacidade de lidar com

dados imperfeitos, denominados ruídos. Muitos conjuntos de dados apresentam esse tipo de caso, sendo alguns erros comuns a presença de dados com rótulos e/ou atributos incorretos. A técnica de Aprendizado de Máquina deve idealmente ser robusta a ruídos presentes nos dados, procurando não fixar a obtenção dos classificadores sobre esse tipo de caso. Deve-se também minimizar a influência de outliers no processo de indução. Os outliers são exemplos muito distintos dos demais presentes no conjunto de dados. Esses dados podem ser ruídos ou casos muito particulares, raramente presentes no domínio.([FACELI; OUTROS, 2017](#))

Devido a necessidade de que a solução desenvolvida possa ser executada no hardware disponível em pequenos estabelecimentos, foram selecionados os seguintes algoritmos clássicos do Aprendizado de Máquina:

- a) *OLS - Ordinal Least Square* (Regressão linear)
- b) *Neural network* (Redes Neurais)
- c) *Support Vector Machine* (SVM Linear e Radial):
- d) *Multi-Layer Perceptron* (MLP)
- e) *Random Forest* (RF)
- f) *k-Nearest Neighbour Regression* (KNN):

O Ensemble de modelos é uma técnica que permite a construção de modelos a partir dos modelos existentes, tem o objetivo de capturar as melhores características de cada um dos modelos utilizados e gerar um novo modelo com melhor acurácia. Neste trabalho onde os modelos são heterogêneos pois os algoritmos representam diferentes técnicas de solução, foram utilizadas as seguintes modos de construção de ensemble:

- **Stacking** ou generalização em pilha ([FACELI; OUTROS, 2017](#)) onde os dados originais são acrescidos dos dados gerados pelos modelos, e em seguida um novo modelo é gerado utilizando todos os dados disponíveis
- **Blending** ou generalização em cascata ([FACELI; OUTROS, 2017](#)) onde os dados gerados pelos modelos são submetidos a um operador construtivo que cria novos atributos, este conjunto de novos atributos é utilizado para a construção de um novo modelo.

6.2 Motor de Modelos

Para atender a necessidade de se construir modelos flexíveis, que atendam a diferentes necessidades do negócio, foram criados processos parametrizados que permitem a manipulação dos seguintes aspectos: Dimensão Temporal, Dimensão Produto, Remoção de variáveis, Variável Preditora, Lista de Algoritmos, Ensemble Stacking e Ensemble Blending.

Este conjunto de processos que constituem um motor de modelos é composto por rotinas codificadas em linguagem R que atendem as seguintes necessidades computacionais:

1. Configuração

Efetua a leitura dos dados disponibilizados pelo processo de Engenharia de Dados, identifica datas (mínima, máxima) associados aos dados, variáveis a serem excluídas do modelo e respectivas classes de dados dos atributos.

2. Pré-processamento dos dados

O processo de engenharia de dados efetuou uma série de transformações nos dados originais para permitir o seu uso na elaboração das estatísticas descritivas, estas transformações resolveram ou identificaram a maioria dos problemas relacionados a ruídos e outliers, porém para o uso no processo de Aprendizado de Máquina é necessário que os dados sejam pré-processados para melhor adequar os dados as particularidades dos algoritmos a serem processados, conforme as especificações recebidas nos parâmetros listados abaixo: (Tabela 11)

Tabela 11 – Parametros da rotina de pré-tratamento dos dados.

Nome	Descrição
df	dados originais
confianca	1
erro	0
idLinha	coluna que contém o valor identificador da linha
colDependente	coluna que contém a variável dependente
dadoQuantitativo	indicador de uso das variáveis quantitativas (dados numéricos)
dadoQualitativo	indicador de uso das variáveis qualitativas (fatores)
remove_NA	indicador de remoção das observações que contêm NA's
ConverteDataHora	indicador de conversão de data/hora para número
centralizar	indicador se os dados devem ser centralizados
escalar	indicador se os dados devem ser escalados
dfAtrib	definição dos atributos, utilizado para excluir colunas dos dados
SepararValorPorFator	Lista de vetores contendo variável de valor
Factor_as_Dummy	vetor de colunas contendo variáveis factor e, m variáveis dummy
Factor_as_Numeric	indicador de conversão de fatores para classe numérico
Remove_nearZeroVar	indicador de exclusão de variáveis com baixa variância
remove_findCorrelation	indicador de exclusão de variáveis altamente correlacionadas
Remove_findLinearCombos	indicador de remoção de variáveis com dependência linear

Fonte: Elaborado pelo autor.

3. Sumarização dos dados

Os dados obtidos da rotina de pré-processamento são submetidos a um processo de sumarização de acordo com as dimensões temporais e de produto desejadas. As dimensões temporais possíveis são: (Tabela 12)

Tabela 12 – Dimensão temporal

Nr. ordem	Código	Descrição	Formato	Frequência
1	ano	dados anuais	9999	2
2	semestre	Dados semestrais (ano dividido em 2 blocos de 6 meses)	99999	6
3	quadrimestre	Dados quadrimestrais (ano dividido em 3 blocos de 4 meses)	99999	4
4	trimestre	Dados trimestrais (ano dividido em 4 blocos de 3 meses)	99999	3
5	bimestre	Dados bimestrais (ano dividido em 6 blocos de 2 meses)	99999	2
6	mes	Dados mensais (ano dividido em 12 meses de acordo com o calendário)	999999	12
7	quinzena	Dados quinzenais (ano dividido em 24 blocos de 15/16 dias de acordo com o calendário dos meses)	9999999	24
8	semana	Dados semanais (ano dividido em 52/53 blocos de 7 dias)	999999	52,25
9	ano_dia	Dados diários (ano dividido em 365 blocos de 1 dia)	9999999	365,25
10	ano_mes_dia	Dados diários (ano dividido em 365 blocos de 1 dia de acordo com o calendário)	99999999	365,25
11	ano_mes_dia_turno	Dados diários por turno (dia dividido em 4 blocos de 6 horas de duração)	99999999	4
12	ano_mes_dia_hora	Dados diários por hora (dia dividido em 24 blocos de 1 hora de duração)	99999999	24
13	ano_mes_dia_minuto	Dados diários por minuto (hora dividido em 60 blocos de 1 minuto de duração)	99999999	60
14	ano_mes_dia_segundo	Dados diários por segundo (hora dividido em 3600 blocos de 1 segundo de duração)	99999999	3600

A escolha da dimensão produto (loja, departamento, seção, grupo e produto) está associada ao objetivo a ser atingido pelo modelo de previsão, questões gerenciais normalmente estão associadas a loja como um todo e questões específicas devem na maioria das vezes serem analisadas pelo nível mais baixo (produto).

4. Avaliação de modelos

Por ultimo os dados sumarizados são utilizados no processamento de diferente algoritmos de Aprendizado de Máquina, a rotina responsável pela execução é controlada pelo parâmetros de entrada que são listados abaixo:

Tabela 13 – Parametros da rotina de avaliação de modelos.

Nome	Descrição
dfDadosModelo	dados de entrada para processamento d modelo
dfAtrib	definição dos atributos, utilizado para excluir colunas dos dados
cQuebra	atributos utilizados para separar os dados em diferentes modelos
cPeriodo	Identificação do periodo: diário, mensal, etc.
cValor	atributo de interesse na predição
datacorteini	Data que separa os dados de treino dos dados para predição
datacortefim	Data que inicia o processamento do ensemble
datafim	Data final para processamento do modelo
lstMetodos	Lista dos métodos que serão executados para os dados
MaxPeriodos	Quantidade máxima de vezes que o modelo será executado
lstExclusao	lista de atributos que devem ser excluídas do modelo
numVidasModelo	numero de vezes que o modelo será utilizado até ser submetido a um novo treinamento

Fonte: Elaborado pelo autor.

Após a identificação dos parâmetros recebidos, os dados são processados de acordo com os algoritmos de Machine Learning especificados em três etapas distintas:

- algoritmo:** Um modelo padrão que utiliza os dados disponibilizados pela engenharia de dados;
- stacking:** modelo que acrescenta aos dados originais os valores obtidos no modelo padrão criando um novo modelo (dados originais + dados de previsão).
- blending:** modelo que utiliza os valores obtidos no modelo padrão para a criação de um novo modelo (dados de previsão).

5. Geração de relatórios e gráficos

Os dados obtidos durante a avaliação de modelos, são recuperados e utilizados para construção de graficos e tabelas que auxiliam na elaboração de conclusões sobre os modelos.

O algoritmo desenvolvido nesta rotina provou ser bastante flexível, pois permite a execução de previsão de vendas, demanda e de preços em diferentes configurações das dimensões temporais e de produto.

O processo de análise utilizou a métrica *RMSE* que é raiz do erro médio quadrático (*Root Mean Square ERROR*), como critério de avaliação dos modelos preditivos. O uso de ensemble modelos busca capturar ganhos de precisão, pois cada algoritmo possui certa superioridade seletiva, sendo o melhor para algumas tarefas mas não para todas (FACELI; OUTROS, 2017).

As variaveis de controle dos diferentes modelos são denominadas de hiperparametros, eles tem seu valor definido antes que o processo de aprendizado comece, alguns modelos não possuem hiperparametros, porém quando existentes eles são primordiais na obtenção de bons resultados. Os hiperparametros utilizados pelo motor de modelos são:

a) *OLS - Ordinal Least Square* (Regressão linear)

Não possui hiperparametros

b) *Neural network* (Redes Neurais)

size = seq(from = 1, to = 10, by = 1),

decay = seq(from = 0.1, to = 0.5, by = 0.1)

c) *Support Vector Machine* (SVM Linear)

cost = c(0, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768)

d) *Support Vector Machine* (SVM Radial)

cost = c(0, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768),

gamma=c(0, 0.5, 1, 1.5, 2.5, 5.0, 10.0, 50.0)

e) *Multi-Layer Perceptron* (MLP)

layer1=5,

layer2=5,

layer3=5

f) *Random Forest* (RF)

mtry=20

g) *k-Nearest Neighbour Regression* (KNN):

k=100

6.3 Previsão de vendas

A previsão de vendas é uma das principais ferramentas para a gestão de uma loja de varejo, para atender às diferentes necessidades do negócio, os dados foram agregados utilizando diferentes combinações das dimensões de tempo e de produto, conforme Tabela 14.

Tabela 14 – Conjuntos de dados

Dimensão temporal	Dimensão produto
Mensal	Total a loja
Mensal	Departamento 1-Drugstore
Mensal	Departamento 2-Perecíveis
Mensal	Departamento 3-Não Perecíveis
Semanal	Total a loja
Semanal	Departamento 1-Drugstore
Semanal	Departamento 2-Perecíveis
Semanal	Departamento 3-Não Perecíveis
Diário	Total a loja
Diário	Departamento 1-Drugstore
Diário	Departamento 2-Perecíveis
Diário	Departamento 3-Não Perecíveis

Fonte: Elaborado pelo autor.

As dimensões temporais atendem a diferentes necessidades do administrador, por exemplo: Mensal (Utilizado para o planejamento gerencial do negócio, permite identificar necessidades de caixa, projeção de lucros, etc), Semanal (Permite identificar de maneira antecipada tendências de desvios na projeção de vendas mensais), Diário (Utilizado como Ferramenta do gerenciamento operacional do negócio).

6.3.1 Mensal

A elaboração de previsões de faturamento, a partir do processamento de todos os dados do SuperX agrupados por mês permitem ao gestor elaborar previsões sobre o fluxo de caixa e estabelecer metas de faturamento a serem perseguidas durante o mês vigente. A análise do RMSE (tabela 41) obtido com o processamento dos dados para toda a loja permite identificar que os modelos Stacking são superiores na qualidade de resposta quando comparados aos demais modelos.

A análise das métricas mostram que os algoritmos com melhores resultados de acordo com os tipos de modelo são:

Tabela 15 – RMSE - Toda a loja com periodicidade mensal.

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	17475.43	2366.48	1240.65	137.45	5030.09	421.89
2 SVM Linear	15321.83	567.79	2404.62	403.27	4994.26	527.40
3 SVM Radial	13249.95	1007.63	9107.80	1131.27	4994.26	396.82
4 Neural Network	13864.17	1393.48	13388.70	1338.00	7436.32	987.92
5 Multi Layer Perceptron	19391.23	2377.18	20150.37	2439.03	28102.01	21609.52
6 Random Forest	11378.55	674.01	8035.76	462.01	7608.22	989.96
7 KNN	16129.55	820.10	11132.84	924.03	6961.58	319.11

Fonte: Elaborado pelo autor

- a) **Padrão:** Randon Forest, SVM Radial e Neural Network;
- b) **Stacking:** OLS (Ordinal Least Squere), SVM Linear e Random Forest.
- c) **Blending:** SVM Linear, SVM Radial e Random Forest.

Destaca-se o algoritmo Random Forest, pela consistência dos dados em todos os três cenários, e de forma negativa o MLP por apresentar os piores resultados em todos os cenários.

A analise dos gráfico exibidos nas Figura 37, 38 e 39, mostram o comportamento do RMSE dos melhores modelos. Os gráficos reforçam a percepção de que os modelos ensemble possuem resultados sensivelmente superiores, e está de acordo com (FACELI; OUTROS, 2017) que cita entre os benefícios dos modelos ensemble o ganho de precisão.

Figura 37 – RMSE da Previsão de vendas mensal: Melhores algoritmos

Fonte: Elaborado pelo autor

Figura 38 – RMSE da Previsão de vendas mensal: Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 39 – RMSE da Previsão de vendas mensal: Melhores Modelos Blending

Fonte: Elaborado pelo autor

Nas tabelas 16, 17 e 18 temos o resultado do processamento dos dados utilizando como filtro o departamento associado ao produto. .

Tabela 16 – RMSE - Drugstore com periodicidade mensal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	601.00	52.02	53.14	11.53	189.75	21.17
2	SVM Linear	533.64	9.24	70.21	13.46	196.65	23.42
3	SVM Radial	483.94	29.11	294.28	25.87	196.65	13.23
4	Neural Network	560.88	10.51	383.65	70.97	232.63	15.45
5	Multi Layer Perceptron	677.04	87.93	610.57	48.80	681.43	60.61
6	Random Forest	418.38	22.73	273.46	42.92	351.95	18.32
7	KNN	452.98	35.55	453.07	50.10	271.00	20.20

Fonte: Elaborado pelo autor

Tabela 17 – RMSE - Não pereceíveis com periodicidade mensal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	7599.75	956.84	746.81	67.36	2381.41	271.80
2	SVM Linear	7238.15	514.98	1221.31	169.64	2799.94	688.38
3	SVM Radial	6906.01	667.33	4163.02	663.03	2799.94	122.60
4	Neural Network	8588.36	474.25	7337.36	1164.32	3691.75	331.03
5	Multi Layer Perceptron	15158.92	9130.31	14552.25	9019.24	14524.71	9410.48
6	Random Forest	5599.42	493.62	4460.60	828.81	5321.92	481.62
7	KNN	6453.38	419.29	5906.51	297.01	4322.80	312.92

Fonte: Elaborado pelo autor

Tabela 18 – RMSE - Pereceíveis com periodicidade mensal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	9141.11	943.49	799.94	123.67	2240.58	217.76
2	SVM Linear	9959.38	1094.72	1346.04	316.93	2620.85	415.12
3	SVM Radial	7184.62	597.52	5162.65	551.24	2620.85	201.76
4	Neural Network	9231.37	569.37	7481.98	808.29	3513.43	300.62
5	Multi Layer Perceptron	11325.88	1388.89	11063.28	1066.46	14811.82	11492.20
6	Random Forest	6405.58	878.43	4515.42	298.35	4562.71	321.01
7	KNN	7151.68	422.44	5645.51	421.27	3874.08	169.93

Fonte: Elaborado pelo autor

Destaca-se o algoritmo MLP, por apresentar a pior performance em todos os cenários. A análise da métrica RMSE mostra que nos modelos ensemble Stacking os algoritmos com melhores resultados são:

a) **Drugstore:** OLS (Ordinal Least Square);

b) **Não Perecíveis:** OLS (Ordinal Least Square);

c) **Perecíveis:** OLS (Ordinal Least Square).

Este procedimento de dividir os dados segundo um critério pré-definido, neste caso de departamentos do SuperX, auxilia o gestor a identificar eventuais discrepâncias no comportamento de vendas. A análise dos gráfico exibidos nas Figura 40, 41 e 42, mostram o comportamento do RMSE nos diferentes cenários. Os algoritmos OLS (Ordinal Least Square) e SVM Linear estão presentes para todos os departamentos e apresentam um comportamento bastante semelhante.

Figura 40 – RMSE da Previsão de vendas mensal (Drugstore): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 41 – RMSE da Previsão de vendas mensal (Não Perecíveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 42 – RMSE da Previsão de vendas mensal (Perecíveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Os administradores de um empreendimento como o SuperX tem uma expectativa de faturamento baseados em dados empíricos e sujeito a avaliações subjetivas. A partir da análise dos resultados dos modelos propostos é possível fornecer previsões de venda com periodicidade mensal e com critérios de acurácia definidos.

O APE (Absolute Percentage Error) para os melhores modelos utilizando stacking é exibido nos gráficos abaixo:

Figura 43 – (%) de desvio absoluto da Previsão de vendas mensal (Drugstore)

Fonte: Elaborado pelo autor

Figura 44 – (%) de desvio absoluto da Previsão de vendas mensal (Não Perecíveis)

Fonte: Elaborado pelo autor

Figura 45 – (%) de desvio absoluto da Previsão de vendas mensal (Perecíveis)

Fonte: Elaborado pelo autor

6.3.2 Semanal

A análise do RMSE obtido com o processamento dos dados do SuperX englobando todos os dados da loja e com periodicidade semanal é exibida na tabela 19.

Tabela 19 – RMSE - Toda a loja com periodicidade semanal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	7940.97	15342.54	728.05	535.00	1084.74	84.65
2	SVM Linear	3312.75	263.56	356.26	24.98	1128.91	126.28
3	SVM Radial	2674.61	80.70	1437.86	117.06	1128.91	74.10
4	Neural Network	3048.83	149.58	1310.75	388.32	1327.14	97.16
5	Multi Layer Perceptron	6800.09	5119.87	5395.87	2441.29	5363.20	2486.24
6	Random Forest	2266.48	82.56	1152.18	68.03	2073.83	85.54
7	KNN	2725.82	120.55	1664.38	79.38	1607.30	73.90

Fonte: Elaborado pelo autor

A inspeção das métricas RMSE de cada um dos cenários estudados, mostram que os algoritmos com melhores resultados de acordo com os tipos de modelo são:

- a) **Padrão:** Randon Forest, SVM Radial e KNN;
- b) **Stacking:** SVM Linear, OLS (Ordinal Least Squere) e Randon Forest.
- c) **Blending:** OLS (Ordinal Least Squere), SVM Linear e SVM Radial.

A análise dos gráficos exibidos nas Figura 46, 47 e 48, mantém a percepção de que os modelos ensemble possuem resultados sensivelmente superiores.

Figura 46 – RMSE da Previsão de vendas semanais: Melhores algoritmos

Fonte: Elaborado pelo autor

Figura 47 – RMSE da Previsão de vendas semanais: Melhores modelos Stacking

Fonte: Elaborado pelo autor

Figura 48 – RMSE da Previsão de vendas semanais: Melhores modelos Blending

Fonte: Elaborado pelo autor

Porém no modelo ensemble Stacking do algoritmo OLS (Ordinal Least Square) é bastante visível que embora ele esteja entre os melhores resultados, ele possui um comportamento longe do ideal, com muitos picos e vales (variações de amplitude) e uma tendência de crescimento do RMSE que indica uma piora da qualidade do algoritmo ao longo do tempo. (figura 48)

O resultado dos do processamento dos dados particionados por departamento, estão exibidos nas tabelas 20, 21 e 22.

Tabela 20 – RMSE - Drugstore com periodicidade semanal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	243.58	210.68	23.85	14.61	81.98	2.62
2	SVM Linear	170.32	11.54	18.95	7.94	83.25	3.93
3	SVM Radial	155.99	3.09	59.54	5.77	83.25	2.06
4	Neural Network	161.25	3.32	57.33	16.75	85.98	3.12
5	Multi Layer Perceptron	207.16	21.98	220.94	54.59	217.60	52.40
6	Random Forest	152.00	5.44	43.06	3.58	143.13	5.09
7	KNN	148.48	3.41	61.53	3.52	103.63	3.44

Fonte: Elaborado pelo autor

Tabela 21 – RMSE - Não pereceíveis com periodicidade semanal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	2927.55	4845.34	242.42	139.65	621.59	63.21
2	SVM Linear	1723.68	81.76	197.05	12.65	634.31	71.29
3	SVM Radial	1511.82	49.83	754.43	101.93	634.31	30.28
4	Neural Network	1759.67	75.49	793.21	181.74	745.51	64.72
5	Multi Layer Perceptron	3189.88	1477.11	3393.60	1963.64	2851.78	349.23
6	Random Forest	1268.13	47.28	596.48	38.68	1264.27	51.33
7	KNN	1558.46	56.71	642.66	34.01	944.01	62.80

Fonte: Elaborado pelo autor

Tabela 22 – RMSE - Pereceíveis com periodicidade semanal.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	3284.17	4524.51	554.59	421.26	545.64	37.91
2	SVM Linear	1802.74	133.18	197.91	13.32	569.52	39.78
3	SVM Radial	1342.26	49.96	792.44	51.05	569.52	51.84
4	Neural Network	1811.37	185.65	718.95	196.12	672.01	55.15
5	Multi Layer Perceptron	3195.03	1785.44	3198.72	1750.66	3251.41	1732.91
6	Random Forest	1198.06	47.74	613.44	51.21	1048.56	41.41
7	KNN	1395.64	60.86	694.26	54.24	864.56	35.31

Fonte: Elaborado pelo autor

A análise das métricas mostram que nos modelos ensemble Stacking, os algoritmos com melhores resultados são:

- a) **Drugstore:** Randon Forest;
- b) **Não Perecíveis:** SVM Linear;
- c) **Perecíveis:** OLS (Ordinal Least Square).

Nos gráficos representados pelas Figura 49, 50 e 51 temos o comportamento dos melhores modelos para o cenário ensemble Stacking.

É possível observar que o comportamento que o algoritmo OLS (Ordinal Least Square) repete o comportamento de grande variação no valor do RMSE, e que o algoritmo SVM Linear tem o comportamento contrário (baixa variação, exceto por uma única ocorrência no departamento de Drugstore).

Figura 49 – RMSE Previsão de vendas semanais (Drugstore): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 50 – RMSE Previsão de vendas semanais (Departamento Não Perecíveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 51 – RMSE Previsão de vendas semanais (Perciveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando stacking é exibido nos gráficos abaixo:

Figura 52 – (%) de desvio absoluto da Previsão de vendas semanais (Drugstore)

Fonte: Elaborado pelo autor

Figura 53 – (%) de desvio absoluto da Previsão de vendas semanais (Departamento Não Perciveis)

Fonte: Elaborado pelo autor

Figura 54 – (%) de desvio absoluto da Previsão de vendas semanais (Perciveis)

Fonte: Elaborado pelo autor

6.3.3 Diário

Os dados de venda com uma periodicidade diária são utilizadas no estabelecimento de metas e pontos de controle de curta duração. Principalmente no estabelecimento de metas de faturamento a serem perseguidas durante o dia vigente e a para a identificação com antecedência de graves desvios no previsto X realizado.

A análise do RMSE obtido com o processamento dos dados para toda a loja e com periodicidade diária é exibida na tabela 23.

Tabela 23 – RMSE - Toda a loja com periodicidade diária.

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	781.43	5.26	84.73	1.40	348.17	32.11
2 SVM Linear	788.87	13.04	92.35	9.29	374.24	98.84
3 SVM Radial	692.47	7.27	294.86	18.04	374.24	4.76
4 Neural Network	936.57	33.90	725.37	124.08	375.21	33.38
5 Multi Layer Perceptron	1264.65	423.93	1232.53	451.24	1183.97	319.12
6 Random Forest	660.19	13.66	129.18	17.34	624.71	12.95
7 KNN	868.39	5.04	267.10	14.72	430.37	38.13

Fonte: Elaborado pelo autor

A análise das métricas mostram que os algoritmos com melhores resultados de acordo com os tipos de modelo são:

- Padrão:** Randon Forest, SVM Radial ee OLS (Ordinal Least Squere);
- Stacking:** OLS (Ordinal Least Squere), SVM Linear e Randon Forest.
- Blending:** OLS (Ordinal Least Squere), SVM Linear e SVM Radial.

Destaca-se que os algoritmos SVM Linear e OLS (Ordinal Least Squere), possuem um desempenho bom em ambos tipos de modelos, e como um aspecto negativo a baixa adequação do algoritmo Multi Layer Perceptron aos dados, apresentando o pior resultado em todos os cenários.

A análise dos gráficos exibidos nas Figura 55, 56 e 57, mantém a percepção de que os modelos ensemble possuem resultados sensivelmente superiores.

Figura 55 – RMSE da previsão de vendas diárias: Melhores algoritmos

Fonte: Elaborado pelo autor

Figura 56 – RMSE da previsão de vendas diárias: Melhores modelos Stacking

Fonte: Elaborado pelo autor

Figura 57 – RMSE da previsão de vendas diárias: Melhores modelos Blending

Fonte: Elaborado pelo autor

A semelhança do executado para a análise em outras dimensões temporais, os dados forma segmentados por departamento, cujos resultados de processamento estão exibidos nas tabelas 24, 25 e 26.

Tabela 24 – RMSE - Drugstore com periodicidade diária.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	46.08	0.18	4.75	0.04	27.00	0.97
2	SVM Linear	46.42	0.23	5.02	0.36	27.87	2.56
3	SVM Radial	44.43	0.28	10.27	0.45	27.87	0.71
4	Neural Network	47.93	0.46	33.51	4.27	27.13	0.98
5	Multi Layer Perceptron	62.45	12.32	64.86	11.78	65.66	15.74
6	Randon Forest	43.70	0.33	4.81	0.68	41.95	0.64
7	KNN	47.04	0.19	21.06	0.49	31.82	1.20

Fonte: Elaborado pelo autor

Tabela 25 – RMSE - Não percebíveis com periodicidade diária.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	434.42	2.67	49.69	1.12	190.09	15.46
2	SVM Linear	438.26	2.78	54.48	6.63	200.34	26.90
3	SVM Radial	395.25	3.84	187.79	9.04	200.34	3.01
4	Neural Network	528.64	24.28	432.81	57.21	207.96	17.34
5	Multi Layer Perceptron	702.67	168.33	681.58	131.37	686.43	144.57
6	Randon Forest	371.46	10.01	90.30	14.40	352.51	10.35
7	KNN	501.16	4.14	172.03	8.67	246.71	18.34

Fonte: Elaborado pelo autor

Tabela 26 – RMSE - Percebíveis com periodicidade diária.

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	403.65	2.12	40.36	0.39	188.13	13.52
2	SVM Linear	408.79	18.38	46.06	8.10	195.55	18.34
3	SVM Radial	352.50	3.46	110.14	5.84	195.55	3.14
4	Neural Network	493.03	19.99	405.43	50.45	190.81	13.41
5	Multi Layer Perceptron	605.33	99.77	593.28	78.61	621.40	164.45
6	Randon Forest	340.19	4.91	51.86	4.84	317.95	4.81
7	KNN	402.66	2.32	101.71	4.04	224.19	12.97

Fonte: Elaborado pelo autor

A análise das métricas mostram que nos modelos ensemble Stacking os algoritmos com melhores resultados são:

- a) **Drugstore:** Randon Forest;
- b) **Não Percebíveis:** OLS (Ordinal Least Square);
- c) **Percebíveis:** OLS (Ordinal Least Square).

Em coerência com as observações anteriores o Multi Layer Perceptron (MLP) foi o algoritmo menos eficiente e o OLS (Ordinal Least Square) destacou-se positivamente no cenário de modelos ensemble Stacking.

Nos gráficos representados pelas Figura 58, 59 e 60 temos o detalhamento do comportamento do RMSE para o cenário ensemble Stacking.

Figura 58 – RMSE da Previsão de vendas semanais (Drugstore): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 59 – RMSE da Previsão de vendas semanais (Não Perciveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 60 – RMSE da Previsão de vendas semanais (Perciveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando stacking é exibido nos gráficos abaixo:

Figura 61 – (%) de desvio absoluto da Previsão de vendas semanais (Drugstore): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 62 – (%) de desvio absoluto da Previsão de vendas semanais (Não Percíveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

Figura 63 – (%) de desvio absoluto da Previsão de vendas semanais (Percíveis): Melhores Modelos Stacking

Fonte: Elaborado pelo autor

6.4 Previsão de demanda

A previsão de demanda é outra ferramentas necessária para a gestão de uma loja de varejo. Foram selecionadas as mesmas dimensões temporais Mensal, Semanal e Diária, já utilizadas na Previsão de vendas, e para minimizar o tempo total de processamento optou-se por limitar a aplicação dos modelos ao primeiros 3 produtos em valor de faturamento. Foram mantidos os algoritmos e as métricas já utilizadas no processamento da Previsão de Vendas.

6.4.1 Produto: PÃO FRANCES

Em conversa com os responsáveis pelo SuperX, foi identificado que o produto PÃO FRANCES sofreu grande flutuação nas quantidades vendidas em função de problemas de abastecimento (inclusive troca de fornecedor) e com o equipamento (quebra de forno).

A análise das tabelas 27, 28 e 29 mostram que os modelos ensemble Blending, possuem uma melhor performance quanta analisada a métrica RMSE.

Tabela 27 – RMSE - PÃO FRANCES (periodicidade:mensal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	4329.71	1846.63	4386.32	2048.99	392.76	217.49
2	SVM Linear	4146.94	2184.41	3968.11	1874.27	362.33	196.01
3	SVM Radial	3465.66	1742.43	3265.99	1555.36	362.33	1974.58
4	Neural Network	3733.32	1795.79	3038.28	1431.97	830.83	460.64
5	Multi Layer Perceptron	4462.83	2326.65	4386.78	2234.04	5752.43	3676.94
6	Random Forest	1072.41	669.52	1253.11	859.79	1463.21	662.81
7	KNN	2386.45	919.19	2826.62	1221.71	2099.30	864.69

Fonte: Elaborado pelo autor

Tabela 28 – RMSE - PÃO FRANCES (periodicidade:semanal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	1113.83	1034.56	1470.11	1212.55	243.64	695.47
2	SVM Linear	1062.13	362.00	1094.35	393.65	107.16	65.92
3	SVM Radial	568.66	129.90	642.05	238.13	107.16	394.13
4	Neural Network	899.38	325.48	316.96	149.13	179.68	82.78
5	Multi Layer Perceptron	1182.40	474.22	1302.42	684.35	1185.56	481.78
6	Random Forest	184.68	65.25	205.95	64.71	188.29	90.91
7	KNN	223.15	80.08	285.49	88.46	202.58	102.74

Fonte: Elaborado pelo autor

Tabela 29 – RMSE - PÃO FRANCES (periodicidade:diária)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	128.77	30.79	128.83	30.74	15.28	4.19
2	SVM Linear	159.90	46.18	160.39	46.46	21.69	22.36
3	SVM Radial	68.81	9.06	62.19	7.98	21.69	49.46
4	Neural Network	143.27	43.10	106.79	37.33	22.93	6.52
5	Multi Layer Perceptron	186.71	77.37	192.80	86.94	188.23	81.57
6	Random Forest	32.40	7.55	32.81	7.06	31.53	6.89
7	KNN	33.61	7.21	33.67	7.02	21.25	3.73

Fonte: Elaborado pelo autor

Nas tabelas acima identifica-se que os algoritmos com melhor performance no cenário ensemble Blending são:

- Mensal:** SVM Linear, SVM Radial e OLS (Ordinal Least Square);
- Semanal:** SVM Linear, SVM Radial e Neural Network;
- Diário:** OLS (Ordinal Least Square), KNN e SVM Linear.

A análise dos gráfico exibidos nas Figura 64, 65 e 66, mostram o comportamento do RMSE para os melhores modelos no cenário ensemble Blending.

Figura 64 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 65 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 66 – RMSE Previsão de demanda - PÃO FRANCES (periodicidade:diária)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 67 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 68 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 69 – (%) de desvio absoluto da Previsão de demanda - PÃO FRANCES (periodicidade:diária)

Fonte: Elaborado pelo autor

6.4.2 Produto: REFR COCA-COLA RETORNÁVEL

O produto REFR COCA-COLA RETORNÁVEL é considerado um chamariz, a falta dele no estoque provoca perdas em vendas segundo os responsáveis pelo SuperX. Porém apresenta margens de lucro muito baixas, ocupa muito espaço nas gondolas e estoques.

A análise das tabelas 30, 31 e 32 mostram que os modelos ensemble Blending, possuem uma melhor performance quando analisada a métrica RMSE.

Tabela 30 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	290.09	27.56	342.44	89.36	82.63	10.56
2	SVM Linear	280.34	19.00	41574052.50	124721339.34	87.04	10.03
3	SVM Radial	226.76	15.41	256.70	24.54	87.04	5.82
4	Neural Network	281.97	24.47	304.70	26.08	112.44	13.51
5	Multi Layer Perceptron	361.51	29.49	389.76	66.24	430.95	77.13
6	Randon Forest	189.35	23.39	178.47	14.27	165.91	22.84
7	KNN	227.34	17.49	207.15	32.11	154.28	11.34

Fonte: Elaborado pelo autor

Tabela 31 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	129.10	281.66	220.27	231.51	20.69	2.83
2	SVM Linear	63.56	4.00	65.87	21.07	21.23	3.17
3	SVM Radial	46.01	1.86	46.00	1.58	21.23	1.20
4	Neural Network	66.58	4.25	58.22	2.71	23.98	2.58
5	Multi Layer Perceptron	104.27	17.46	109.23	21.37	105.62	18.38
6	Randon Forest	44.93	2.29	43.32	2.04	34.98	1.30
7	KNN	48.72	1.95	46.06	1.82	30.63	2.06

Fonte: Elaborado pelo autor

Tabela 32 – RMSE - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	13.35	0.23	13.35	0.23	7.12	0.68
2	SVM Linear	13.58	0.93	13.47	0.48	7.82	2.45
3	SVM Radial	12.28	0.25	12.25	0.25	7.82	0.46
4	Neural Network	16.27	0.81	16.23	0.76	7.26	0.64
5	Multi Layer Perceptron	19.65	2.07	19.52	1.72	19.55	1.89
6	Randon Forest	11.84	0.29	11.81	0.27	11.36	0.28
7	KNN	15.70	0.28	15.69	0.28	8.41	0.60

Fonte: Elaborado pelo autor

A análise das métricas obtidas nos diferentes períodos processados, mostram que nos modelos híbridos os algoritmos com melhores resultados são:

- a) **Mensal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- b) **Semanal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- c) **Diário:** OLS (Ordinal Least Square), SVM Linear e KNN.

A análise dos gráfico exibidos nas Figura 70, 71 e 72, mostram o comportamento do RMSE no cenário ensemble Blending.

Figura 70 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 71 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 72 – RMSE Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 73 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 74 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 75 – (%) de desvio absoluto da Previsão de demanda - REFR COCA-COLA RETORNÁVEL (periodicidade:diária)

Fonte: Elaborado pelo autor

6.4.3 Produto: MUSSARELA

A alta demanda do produto MUSSARELA em alguns períodos foi explicada pelos responsáveis do SuperX como sendo resultado de um período de vendas promocionais, normalmente ele possui uma demanda uniforme associada a venda de produtos matinais.

A análise das tabelas 33, 34 e 35 mostram que para o produto Mussarela a classificação dos modelos ensemble Blending, como possuidores de uma melhor performance quanta analisada a métrica RMSE.

Tabela 33 – RMSE - MUSSARELA (periodicidade:mensal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	18.56	1.71	17.52	3.16	4.75	0.51
2	SVM Linear	18.67	1.75	19.08	1.48	5.22	0.58
3	SVM Radial	15.14	1.07	13.57	0.70	5.22	0.59
4	Neural Network	17.53	1.08	17.80	1.46	5.60	0.70
5	Multi Layer Perceptron	25.23	5.20	25.32	3.98	27.78	5.31
6	Random Forest	13.08	1.16	12.82	1.60	8.27	0.50
7	KNN	14.84	0.80	13.14	1.22	7.79	0.79

Fonte: Elaborado pelo autor

Tabela 34 – RMSE - MUSSARELA (periodicidade:semanal)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	6.43	4.67	7.69	5.59	1.57	0.17
2	SVM Linear	4.85	2.71	10.55	8.31	1.61	0.19
3	SVM Radial	3.72	0.09	3.90	0.31	1.61	0.20
4	Neural Network	4.53	0.16	3.73	0.18	1.66	0.16
5	Multi Layer Perceptron	5.69	0.45	5.75	0.55	5.77	0.49
6	Random Forest	3.43	0.12	3.41	0.10	3.06	0.09
7	KNN	3.47	0.10	3.44	0.09	2.38	0.15

Fonte: Elaborado pelo autor

Tabela 35 – RMSE - MUSSARELA (periodicidade:diária)

	algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1	OLS (Ordinal Least Square)	1.20	0.01	1.20	0.01	0.66	0.04
2	SVM Linear	1.22	0.04	1.22	0.05	0.69	0.10
3	SVM Radial	1.16	0.01	1.15	0.01	0.69	0.03
4	Neural Network	1.38	0.08	1.38	0.07	0.63	0.04
5	Multi Layer Perceptron	1.71	0.08	1.72	0.08	1.71	0.08
6	Random Forest	1.13	0.01	1.13	0.01	1.09	0.01
7	KNN	1.55	0.02	1.55	0.02	0.76	0.04

Fonte: Elaborado pelo autor

No cenário ensemble Blending os melhores algoritmos por dimensão temporal são:

- a) **Mensal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- b) **Semanal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- c) **Diário:** OLS (Ordinal Least Square), SVM Linear e KNN.

A análise dos gráfico exibidos nas Figura 76, 77 e 78, mostram o comportamento dos valores do RMSE para os melhores modelos do cenário ensemble Blending.

Figura 76 – RMSE Previsão de demanda - MUSSARELA (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 77 – RMSE Previsão de demanda - MUSSARELA (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 78 – RMSE Previsão de demanda - MUSSARELA (periodicidade:diária)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 79 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 80 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:semanal)

Fonte: Elaborado pelo autor

Figura 81 – (%) de desvio absoluto da Previsão de demanda - MUSSARELA (periodicidade:diária)

Fonte: Elaborado pelo autor

6.5 Previsão de preço

Uma ferramenta de previsão de preço fornece ao administrador elementos para identificar oportunidades de ganhos com a formação de estoques especulativos, permitindo ao administrador capturar oportunidades de elevação de margem em produtos que têm um comportamento de preços sazonal. Ela ajuda a estabelecer pontos de alerta quando os preços praticados estão muito distantes do preços previstos.

6.5.1 Produto: FRANGO FILET

Em termos de faturamento é o 5º produto em valor de vendas, porém não tem um bom acompanhamento quanto a precificação, o valor de venda é estabelecido em função do preço de compra.

A análise das tabelas 36 e 37 que contém o RMSE obtido com o processamento dos dados nas periodicidades mensal e semanal permite identificar que a semelhança da previsão de demanda o cenário ensemble Blending apresenta os algoritmos com melhor performance na métrica RMSE.

Tabela 36 – RMSE - Previsão de preço - FRANGO FILET (periodicidade: mensal)

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	0.29	0.12	0.30	0.10	0.08	0.06
2 SVM Linear	0.29	0.12	0.29	0.16	0.11	0.03
3 SVM Radial	0.27	0.13	0.26	0.11	0.11	0.08
4 Neural Network	0.19	0.14	0.24	0.11	0.09	0.03
5 Multi Layer Perceptron	0.64	0.05	0.63	0.05	0.60	0.03
6 Random Forest	0.21	0.09	0.19	0.07	0.16	0.09
7 KNN	0.25	0.11	0.22	0.10	0.12	0.05

Fonte: Elaborado pelo autor

Tabela 37 – RMSE - Previsão de preço - FRANGO FILET (periodicidade: semanal)

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	0.42	0.41	0.52	0.48	0.04	0.01
2 SVM Linear	0.33	0.13	0.52	1.19	0.08	0.01
3 SVM Radial	0.26	0.10	0.27	0.09	0.08	0.05
4 Neural Network	0.19	0.07	0.15	0.05	0.05	0.01
5 Multi Layer Perceptron	0.62	0.08	0.64	0.08	0.63	0.09
6 Random Forest	0.15	0.05	0.14	0.06	0.15	0.04
7 KNN	0.19	0.06	0.17	0.05	0.10	0.03

Fonte: Elaborado pelo autor

Dos quadros acima extrai-se que os melhores algoritmos são:

- a) **Mensal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- b) **Semanal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;

A análise dos gráfico exibidos nas Figura 82 e 83, mostram o comportamento da métrica RMSE para o cenário ensemble Blending.

Figura 82 – RMSE Previsão de preço - FRANGO FILET (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 83 – RMSE Previsão de preço - FRANGO FILET (periodicidade:semanal)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 84 – (%) de desvio absoluto da Previsão de preço - FRANGO FILET (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 85 – (%) de desvio absoluto da Previsão de preço - FRANGO FILET (periodicidade:semanal)

Fonte: Elaborado pelo autor

6.5.2 Produto: BATATA LAVADA

Em termos de faturamento é o 22º produto em valor de vendas, a precificação deste produto também é estabelecida em função do preço de compra. A análise das tabelas 38 e 39 que contém o RMSE obtido com o processamento dos dados nas periodicidades mensal e semanal identifica o cenário ensemble Blending como gerador de melhores resultados.

Tabela 38 – RMSE - Previsão de preço - BATATA LAVADA (periodicidade: mensal)

algoritmo	Modelo Padrão - RMSE Médio	Modelo Padrão - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	1.07	0.17	1.09	0.15	0.39	0.07
2 SVM Linear	1.01	0.06	1.14	0.16	0.41	0.08
3 SVM Radial	1.03	0.09	0.99	0.07	0.41	0.06
4 Neural Network	0.89	0.04	0.94	0.09	0.44	0.08
5 Multi Layer Perceptron	1.20	0.08	1.20	0.07	0.92	0.21
6 Random Forest	0.95	0.11	0.89	0.07	0.74	0.08
7 KNN	1.07	0.07	1.06	0.07	0.65	0.05

Fonte: Elaborado pelo autor

Tabela 39 – RMSE - Previsão de preço - BATATA LAVADA (periodicidade: semanal)

algoritmo	Modelo Padrão - RMSE Médio	Modelo Padrão - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	1.14	0.68	1.30	0.73	0.19	0.02
2 SVM Linear	0.96	0.04	1.13	0.47	0.19	0.02
3 SVM Radial	0.92	0.04	0.93	0.03	0.19	0.07
4 Neural Network	0.91	0.04	0.86	0.04	0.20	0.03
5 Multi Layer Perceptron	1.31	0.08	1.31	0.06	1.16	0.20
6 Random Forest	0.68	0.05	0.67	0.04	0.56	0.05
7 KNN	0.77	0.04	0.76	0.03	0.40	0.05

Fonte: Elaborado pelo autor

Os algoritmos com melhor resultado no cenário ensemble Blending são:

- a) **Mensal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- b) **Semanal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;

A análise dos gráfico exibidos nas Figura 86 e 87, mostram o comportamento da métrica RMSE para o cenário ensemble Blending.

Figura 86 – RMSE Previsão de preço - BATATA LAVADA (periodicidade: mensal)

Fonte: Elaborado pelo autor

Figura 87 – RMSE Previsão de preço - BATATA LAVADA (periodicidade:semanal)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 88 – (%) de desvio absoluto da Previsão de preço - BATATA LAVADA (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 89 – (%) de desvio absoluto da Previsão de preço - BATATA LAVADA (periodicidade:semanal)

Fonte: Elaborado pelo autor

6.5.3 Produto: BANANA NANICA

Em termos de faturamento é o 39º produto em valor de vendas, e como os demais produtos analisados existe uma deficiência gerencial na atividade de precificação. A análise das tabelas 40 e 41 permite identificar que o comportamento do produto BANANA NANICA é semelhante aos demais, repetindo uma melhor performance para o cenário ensemble Blending.

Tabela 40 – RMSE - Previsão de preço - BANANA NANICA (periodicidade:mensal)

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	0.66	0.06	0.65	0.09	0.15	0.03
2 SVM Linear	0.64	0.06	9284546641.74	24461763853.78	0.16	0.02
3 SVM Radial	0.59	0.06	0.56	0.15	0.16	0.06
4 Neural Network	0.36	0.04	0.41	0.08	0.17	0.03
5 Multi Layer Perceptron	0.91	0.04	0.93	0.07	0.53	0.18
6 Random Forest	0.38	0.04	0.39	0.05	0.32	0.07
7 KNN	0.59	0.05	0.49	0.08	0.36	0.05

Fonte: Elaborado pelo autor

Tabela 41 – RMSE - Previsão de preço - BANANA NANICA (periodicidade:semanal)

algoritmo	Modelo Padrao - RMSE Médio	Modelo Padrao - Desvio padrão RMSE	Modelo Stacking - RMSE Médio	Modelo Stacking - Desvio padrão RMSE	Modelo Blending - RMSE Médio	Modelo Blending - Desvio padrão RMSE
1 OLS (Ordinal Least Square)	0.76	0.72	0.81	0.51	0.08	0.01
2 SVM Linear	16.68	6.88	36.82	68.78	0.11	0.04
3 SVM Radial	0.48	0.05	0.64	0.03	0.11	0.03
4 Neural Network	0.33	0.05	0.30	0.04	0.09	0.01
5 Multi Layer Perceptron	0.91	0.04	0.91	0.04	0.15	0.04
6 Random Forest	0.23	0.02	0.22	0.02	0.21	0.01
7 KNN	0.28	0.02	0.24	0.03	0.16	0.01

Fonte: Elaborado pelo autor

Os algoritmos com melhor resultado no cenário ensemble Blending são:

- a) **Mensal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;
- b) **Semanal:** OLS (Ordinal Least Square), SVM Linear e Rando Forest;

A análise dos gráfico exibidos nas Figura 90 e 91, mostram o comportamento da metrica RMSE para para os melhores algoritmos no cenário ensemble Blending.

Figura 90 – RMSE Previsão de preço - BANANA NANICA (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 91 – RMSE Previsão de preço - BANANA NANICA (periodicidade:semanal)

Fonte: Elaborado pelo autor

O APE (Absolute Percentage Error) para os melhores modelos utilizando blending é exibido nos gráficos abaixo:

Figura 92 – (%) de desvio absoluto da Previsão de preço - BANANA NANICA (periodicidade:mensal)

Fonte: Elaborado pelo autor

Figura 93 – (%) de desvio absoluto da Previsão de preço - BANANA NANICA (periodicidade:semanal)

Fonte: Elaborado pelo autor

6.6 Observações

O uso do motor de modelos permite explorar os melhores algoritmos para uma determinada configuração do problema. A dimensão temporal (diária, semanal, mensal, etc.), a dimensão do produto (grau de agregação das informações, loja, departamento, seção, produto, etc.) e o objetivo do modelo (predição de valor de venda, quantidade a ser vendida, ou preço de venda) tem impactos de maneira diversa em cada um dos algoritmos.

O uso do ensemble mostrou ser extremamente eficaz em melhorar os resultados da métrica de avaliação dos modelos, ficando claro que o método do ensemble que apresenta os melhores resultados é determinado pela configuração do problema.

Fica caracterizado que o algoritmo MLP (Multi-layer perceptron) possui a pior performance de forma consistente em relação aos demais algoritmos.

Para as atividades de desenvolvimento do motor de modelos foram codificados os seguintes programas:

Tabela 42 – Lista de artefatos *codigo fonte* – *AnaliseMultivariada*

Programa	Descrição resumida
P040_CalculaFrequencia_V1	Executa o pré-processamento dos dados.
P041_DefineAtributos_V1	Cálculo da correlação e regressão linear dos dados.
P042_ResumirDados_V1	Cálculo dos diversos métodos de componentes principais.
P043_SeriesTemporais_V1	Análise de agrupamento.
P044_AvaliaModelos_V1	Cálculo dos diversos métodos de componentes principais.
P045_ProcessarModelos_V1	Análise de agrupamento.

Fonte: Elaborado pelo autor

CONCLUSÃO E TRABALHOS FUTUROS

Os supermercados de pequeno porte sofrem de problemas variados consequentes de uma gestão empírica, devido à falta de conhecimento de novas técnicas (e técnicas nem tão novas assim), limitação de recursos tecnológicos e financeiros, e por ultimo mas não menos importante, a resistência à inovação.

O SuperX, estabelecimento fornecedor dos dados utilizados, retrata de forma geral que os problemas operacionais e financeiros poderiam ser mitigados com o uso de ferramentas matemáticas que podem ajudar em vários aspectos da operação e do planejamento do negócio. O conjunto de métodos e processos utilizados neste trabalho demonstrou a existência de pontos que requerem aprimoramento na gestão do SuperX.

A entrevistas com o proprietário e o gerente identificaram o não uso de todas as potencialidades do software instalado para a gestão do SuperX. O software de ERP, quando devidamente configurado, pode auxiliar e muito na qualidade dos dados. Neste caso, as informações de cunho financeiro (vendas com cartão de crédito, por exemplo), tributário (alíquotas de impostos) e operacional (identificação de operadores e supervisores, entrada de produtos no estoque por compra) não foram devidamente atualizados, impossibilitando de se agregar estes atributos na análise.

O processo de engenharia de dados, detectou a existência de falhas na codificação dos produtos com descrições e/ou classificações incorretas, O cadastramento dos produtos é crítico, a ausência do cadastro ou a baixa qualidade do mesmo podem levar a falsas conclusões nas atividades rotineiras do empreendimento, bem como em estudos de caráter estatístico ou de geração de modelos de previsão.

A estatística descritiva identificou queda do faturamento em valores deflacionados, não cadastramento de novos produtos e o consequente uso do produto "balas" para registrar a venda dos mesmos, foi identificada também a possibilidade de se aumentar o faturamento com a extensão do horário de fechamento do SuperX

A previsão das vendas mostrou que existe uma perda de importância do departamento de não perecíveis em relação ao departamento de perecíveis, este é um fato que deve ser melhor investigado, pois de acordo com o proprietário as margens das vendas do departamento de perecíveis são melhores que as do departamento de não perecíveis, desta forma está identificado que ocorre uma diminuição da margem total das vendas do SuperX, com impactos na própria viabilidade do empreendimento.

A previsão da demanda permitiu a identificação de picos e/ou vales (valores extremos ou outliers), nos dados quantitativos de venda dos produtos analisados demonstrando ser uma ferramenta que indica pontos de atenção a serem avaliados pela gerencia, bem como pratica constante de guiar as compras em função das quantidades previstas iria gerar uma otimização das vendas pelo abastecimento constante das gondolas, minimizando perdas decorrentes de falta de produtos o estabelecimento ou da compra de produtos em excesso. Segundo o proprietário o faturamento poderia crescer 10% com a recuperação das perdas de vendas por falta de produtos nas prateleiras.

A previsão de preços permitiu a identificação de preços congelados por longos períodos indicativo de deficiência na gestão de preços, e demonstrou a possibilidade de ser utilizada como ferramenta para avaliar a formação de estoques especulativos.

Em uma entrevista final com o proprietário e o gerente responsável para discussão dos resultados preliminares deste trabalho, eles se mostraram convencidos ue o grau de envolvimento do gerente/ responsável nas questões operacionais e administrativas, nos diferentes aspectos relacionados ao uso dos dados afeta diretamente os resultados operacionais e financeiros do empreendimento.

Como contribuição para trabalhos futuros que deem continuidade a este projeto, são identificadas as seguintes oportunidades:

- a) **Análise Multivariada:** Uso da análise multivariada, em particular os algoritmos de Componentes Principais e de Agrupamento, na busca de atributos que possam melhorar os modelos de previsão.
- b) **Series Temporais:** Incorporação de modelo de series temporais (ARIMA e SARIMA) entre os algoritmos utilizados nos diferentes modelos de previsão.
- c) **Quebra de gôndola:** Com base no histórico de vendas e da previsão de demanda, identificar os produtos com alta probabilidade de vendas cessantes durante o período de operação do SuperX devido a falta do produto nas gondolas.
- d) **Detectar fraudes:** No setor de açougue as carnes bovinas são resultado do fracionamento de dianteiros e/ou traseiros do boi, sendo conhecidos em termos percentuais a composição de cada corte que pode ser obtido. Neste caso a avaliação em termos percentuais das vendas por corte em um período semanal deve se aproximar deste valores. Quando as vendas não refretem estes percentuais existe a possibilidade

de estarem ocorrendo fraudes no açougue e/ou nos PDV's pela troca de código de um produto nobre (por exemplo: picanha, filet mignon) por outro de menor valor (musculo, acem, etc).

- e) **Cesto de compras:** A correta identificação de quais itens são vendidos em conjunto é uma ferramenta que auxilia o administrador na criação de promoções, e posicionamento dos produtos dentro do espaço do supermercado.
- f) **Planejar produção:** Na padaria os pães devem ser assados preferencialmente várias vezes ao dia para estarem sempre fresquinhos, e vendidos no dia. Esta necessidade leva ao problema de planejar a produção em termos de quantas fornadas, em quais horários e em que quantidade devem ser assados os pães. A produção não deve ultrapassar muito a demanda do dia, pois os pães do dia anterior não são comercializados, porém estes são insumo para a produção de farinha de rosca.
- g) **Planejar pessoal:** O número de PDV's em funcionamento para não formar filas no caixa é obtido a partir da previsão horária de quantidade de vendas (numero de cupons fiscais emitidos por hora) ela permite melhorar o atendimento na loja, gerando satisfação aos clientes, além de identificar horários de ociosidade onde os funcionários podem ser direcionados para trabalho de reposição e organização das gôndolas.
- h) **Planejar compras:** Produtos em promoção no fornecedor podem ser uma oportunidade de melhorar a margem do negócio com a formação de estoques desde que condizentes à previsão de vendas para o período de validade do produto e os custos financeiros. Os produtos com comportamento sazonal de preços (leite, frutas, verduras, produtos natalinos) são beneficiados por uma previsão de demanda podendo identificar oportunidades para formação de estoques, produtos a serem retirados da lista de compras, etc.

De forma adicional um trabalho futuro deveria também: Identificar os motivos que levam os modelos baseados em redes neurais (Neural Network e Perceptron) a apresentarem de forma consistente piores resultados em comparação aos outros modelos, e trabalhar com novos algoritmos em busca de melhores resultados dos modelos preditivos

REFERÊNCIAS

(ABRAS), A. B. de S. **40 anos de Supermercados no Brasil**. [S.l.], 1993. Citado na página 29.

_____. **Ranking dos maiores Supermercadistas**. [S.l.], 2011. Citado na página 30.

ARAÚJO, A. L. V. d. **Aplicação de Regras de Associação para Auxílio na Gestão de Vendas de uma Empresa Varejista Utilizando a ferramenta WEKA**. Dissertação (Mestrado) — Escola Politécnica de Pernambuco – Universidade de Pernambuco, 2009. Citado nas páginas 39, 40, 41, 42 e 43.

BADIN, N. T. **AVALIAÇÃO DA PRODUTIVIDADE DE SUPERMERCADOS E SEU BENCHMARKING**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 1997. Citado nas páginas 11, 27, 28 e 33.

CARVALHO, A. P. L. F. **Redes Neurais Artificiais**. INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO - USP, 2009. Disponível em: <<http://conteudo.icmc.usp.br/pessoas/andre/research/neural/>>. Citado na página 32.

CASTRO, L. N. d.; ZUBEN, F. J. V. **Redes Neurais Artificiais**. Faculdade de Engenharia Elétrica e de Computação - UNICAMP, 2003. Disponível em: <ftp://vm1-dca.fee.unicamp.br/pub/docs/vonzuben/ia006_03/topico5_03.pdf>. Citado na página 32.

CORTES, S. D. C.; PORCARO, R. M. L. **Mineração de Dados - Funcionalidades, Técnicas e Abordagens**. [S.l.: s.n.], 2002. Citado na página 39.

FACELI, K.; OUTROS e. **Inteligência Artificial, Uma abordagem de Aprendizado de Máquina**. [S.l.]: LTC - Livros Técnicos e Científicos Editora Ltda., 2017. Citado nas páginas 31, 32, 33, 85, 86, 89 e 91.

FILHO, U. A. S. **O setor supermercadista no Brasil nos anos 1990**. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz, 2003. Citado na página 29.

GVCEV, C. d. E. e. V. **Relatório ANÁLISE SETORIAL SUPERMERCADOS 2011**. FGV-EAESP, 2012. Disponível em: <https://cev.fgv.br/sites/cev.fgv.br/files/AnaliseSetorial_Supermercados_2011.pdf>. Citado na página 29.

HAN, J.; KAMBER, M. **Data Mining - Concepts and Techniques**. [S.l.]: Morgan Kaufmann Publishers, 2001. Citado na página 39.

IBGE, I. B. d. G. e. E. **Demografia das Empresas**. 2015. Disponível em: <<https://sidra.ibge.gov.br/pesquisa/demografia-das-empresas/tabelas>>. Citado na página 30.

LEVY, M. **Retailing Management**. [S.l.]: McGraw-Hill Education, 2014. Citado nas páginas 28 e 29.

- MILIES, F. C. P.; OUTROS e. **A MATEMÁTICA INTERATIVA NA INTERNET**. Instituto de Matemática e Estatística - USP, 2019. Disponível em: <<http://www.matematica.br/historia/requacoes.html>>. Citado na página 31.
- NETO, F. L.; DINIZ, C. A. R. **Data Mining - Uma introdução**. [S.l.]: São Paulo, 2000. Citado na página 39.
- NETO, P. L. d. O. C. **Estatística**. [S.l.]: Editora Edgard Blucher, 1977. Citado nas páginas 31 e 55.
- PICKOVER, C. A. **LIVRO DA MATEMÁTICA, O - DE PITÁGORAS A 57ª DIMENSÃO**. [S.l.]: librero, 2011. Citado na página 31.
- R-PROJECT, T. **THE R-PROJECT FOR STATISTICAL COMPUTING**. R FOUNDATION, 1997. Disponível em: <<https://www.r-project.org/about.html>>. Citado na página 46.
- REEVES, P. **WHATS HAPPENING WITH FIREBIRD**. IBPHOENIX, 2001. Disponível em: <http://www.ibphoenix.com/resources/documents/search/doc_15>. Citado na página 43.
- SHIKIDA, P. F. A.; MARGARIDO, M. A. **UMA ANÁLISE ECONOMETRICA DE SAZONALIDADE DOS PREÇOS DA CANA-DE-AÇÚCAR, ESTADO DO PARANÁ, 2001-2007**. Instituto de Economia Agrícola (IEA), 2009. Disponível em: <<http://www.iea.sp.gov.br/ftp/iea/publicacoes/IE/2009/tec1-0209.pdf>>. Citado na página 60.
- SILVA, J. C. d. **Aprendendo em uma Floresta Aleatória**. MEDIUM, 2018. Disponível em: <<https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleatoria-3545f6babdf8>>. Citado na página 33.