instituto politécnico de gestão e tecnologia
centro de investigação

Pós Graduação Analytics e Data Science Empresarial

Modelo de Previsão de Vendas Multiproduto

Autor
Ricardo Castro (rc@rcoe.pt)
Orientadores
João Alves (Metyis Porto)
Nelson Pires (ISLA)
Firmino Silva (ISLA)

INTRODUÇÃO

O projeto apresentado, foi realizado em colaboração com a Metyis e destina-se à “The Greenery” uma cooperativa de agricultores holandeses dedicada ao comércio e venda de frutas e vegetais, com origem na Holanda e não só, e destinados à distribuição nacional e internacional, essencialmente para supermercados e grandes superfícies. Mantendo o foco na rentabilidade do negócio e resultados financeiros crescentes, a “The Greenery”, pretende ajustar a oferta à procura reduzindo o desperdício e evitando rotura de stocks.

A operar num mercado cada vez mais exigente e concorrencial, e sentindo a necessidade de otimizar os seus processos pela inovação e digitalização, a “The Greenery”, forneceu via Metyis, os dados da sua atividade, de modo a que aplicando técnicas e modelos de data science e business analytics, lhes fosse desenvolvida uma ferramenta de utilização simples, capaz de fazer previsões de vendas dos vários produtos com que trabalha.

Palavras-chave:
CRISP-DM, SARIMAX, MACHINE LEARNING; PYTHON; THE GREENERY.

OBJETIVO

Desenvolver uma ferramenta de previsão baseada em linguagem Phyton, aplicando a metodologia CRISP-DM e utilizando técnicas de data Science, com modelos matemáticos, estatísticos e de Machine Learning, como algoritmos de classificação, redução de variáveis e modelos autorregressivos integrando variáveis móveis (SARIMAX), com o objetivo último de produzir previsões de venda para os produtos considerados mais relevantes, numa janela temporal de 8 semanas, considerando a influência de fatores exógenos, no caso as condições meteorológicas.

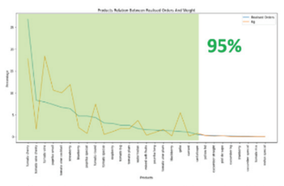
A metodologia seguida neste projeto foi a metodologia CRISP-DM - CRoss Industry Standard Process for Data Mining, nas 6 fases conhecidas:

- Business understanding;
- Data understanding;
- Data Preparation;
- Modeling;
- Evaluation;
- Deployment.

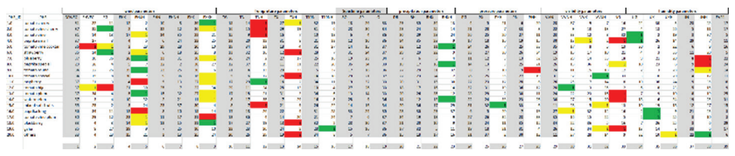
Fluxograma da metodologia CRISP-DM

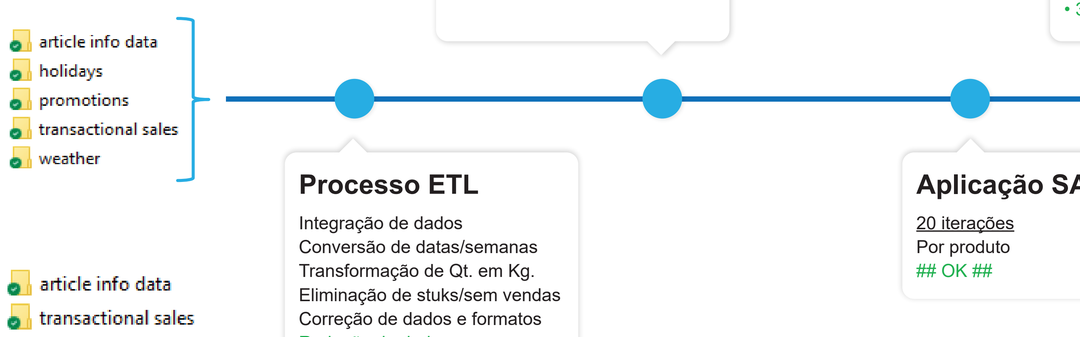
IMPLEMENTAÇÃO DO PROJECTO

A implementação do projeto seguiu várias etapas conforme workflow que se apresenta, passando no essencial pela verificação da qualidade dos dados, ingestão no processo, e tratamento cuidado e criterioso, não só para aproveitar ao máximo a informação recebida, mas também garantir que a mesma apresenta formatos, coerentes, legíveis e integráveis com todo o processo de modelação. A modelação teve por base a utilização de modelos autorregressivos integrando variáveis móveis (SARIMAX), aplicado a séries temporais, integrando a influência das condições climatéricas como variáveis exógenas ao processo de previsão. Os modelos criados seguiram técnicas matemáticas, estatísticas e de Machine Learning aprendidas.



Conjunto de artigos mais vendidos pelo critério definido


Indicadores de condições meteorológicas dominantes para cada série de produtos

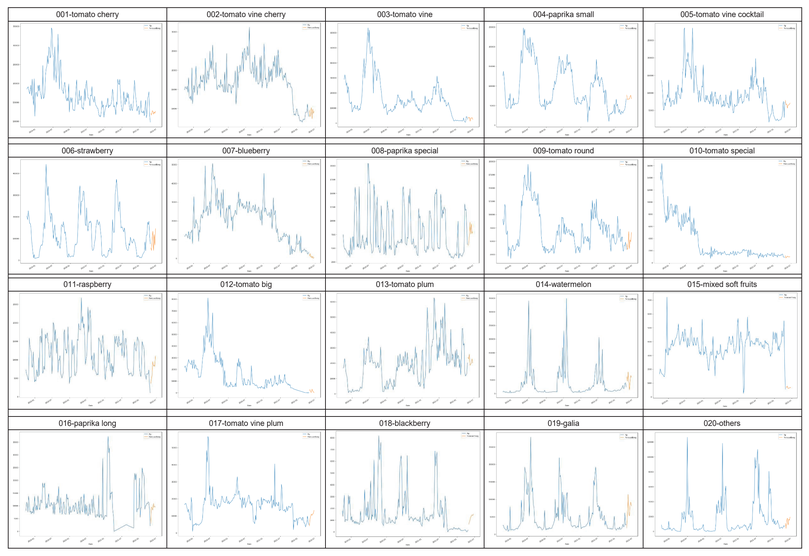


	CodeID	PHID	PH3	Model_CE	Model	p	d	q	P	D	Q	m	AIC_CE	Status
0	000001		1 tomato cherry	SARIMAX(0, 1, 2)	SARIMAX	0	1	2	-	-	-	-	4313,811	OK
1	000002		2 tomato vine cherry	SARIMAX(0, 1, 1)x(1, 0, [1, 52)	SARIMAX	0	1	1	1	0	[1	52	3451,105	OK
2	000003		3 tomato vine	SARIMAX(0, 1, 0)	SARIMAX	0	1	0	-	-	-	-	5236,849	OK
3	000004		4 paprika small	SARIMAX(0, 1, 0)	SARIMAX	0	1	0	-	-	-	-	4943,122	OK
4	000005		5 tomato vine cocktail	SARIMAX(2, 1, 1)x(1, 0, [1, 52)	SARIMAX	2	1	1	1	0	[1	52	4312,627	OK
5	000006		6 strawberry	SARIMAX(1, 0, 1)	SARIMAX	1	0	1	-	-	-	-	4570,776	OK
6	000007		7 blueberry	SARIMAX(2, 1, 0)x(2, 0, 0, 52)	SARIMAX	2	1	0	2	0	0	52	3688,558	OK
7	000008		8 paprika special	SARIMAX(3, 0, 0)x(0, 0, [1], 52)	SARIMAX	3	0	0	0	0	[1]	52	3553,655	OK
8	000009		9 tomato round	SARIMAX(3, 0, 0)x(2, 0, 0, 52)	SARIMAX	3	0	0	2	0	0	52	4204,031	OK
9	000010		10 tomato special	SARIMAX(1, 1, 0)x(0, 0, [1], 52)	SARIMAX	1	1	0	0	0	[1]	52	3039,362	OK
10	000011		11 raspberry	SARIMAX(1, 0, 0)x(1, 0, 0, 52)	SARIMAX	1	0	0	1	0	0	52	3596,414	OK
11	000012		12 tomato big	SARIMAX(2, 1, 2)	SARIMAX	2	1	2	-	-	-	-	3412,911	OK
12	000013		13 tomato plum	SARIMAX(1, 0, 0)	SARIMAX	1	0	0	-	-	-	-	4639,715	OK
13	000014		14 watermelon	SARIMAX(3, 0, 2)x(0, 0, [1], 52)	SARIMAX	3	0	2	0	0	[1]	52	4402,247	OK
14	000015		15 mixed soft fruits	SARIMAX(3, 0, 0)x(0, 0, [1, 2], 52)	SARIMAX	1	0	0	0	0	[1, 2]	52	3031,15	OK
15	000016		16 paprika long	SARIMAX(1, 0, 0)	SARIMAX	1	0	0	-	-	-	-	3164,991	OK
16	000017		17 tomato vine plum	SARIMAX(2, 0, 0)	SARIMAX	2	0	0	-	-	-	-	3723,701	OK
17	000018		18 blackberry	SARIMAX(1, 0, 0)	SARIMAX	1	0	0	-	-	-	-	3146,8	OK
18	000019		19 galia	SARIMAX(2, 0, 0)x(1, 0, 0, 52)	SARIMAX	2	0	0	1	0	0	52	4405,348	OK
19	000020		20 others	SARIMAX(2, 0, 0)	SARIMAX	2	0	0	-	-	-	-	4111,65	OK

Parâmetros SARIMAX obtidos por rotina automática em Python

	PHID	PH3	Week_1	Week_2	Week_3	Week_4	Week_5	Week_6	Week_7	Week_8
0	1	tomato cherry	120790,73	116674,73	127813,14	126706,34	119545,53	124012,49	121375,43	126737,22
1	2	tomato vine cherry	9682,68	7829,32	9004,37	10220,4	7354,01	9837,63	8772,81	9208,42
2	3	tomato vine	39072,89	31955,55	38398,53	12235,26	33302,8	32712,12	36460,1	23767,5
3	4	paprika small	76650,87	65410,57	65508,37	67233,32	66684,25	73319,01	76138,86	67360
4	5	tomato vine cocktail	79211,84	66617,12	67785,92	56546,84	65638,61	65366,27	67171,83	71243,99
5	6	strawberry	53417,5	44840,61	105980,67	113517,2	44568,31	118735,33	69290,05	145655,98
6	7	blueberry	1254,1	2624,28	789,04	1296,08	258,38	1128,17	39,36	431,16
7	8	paprika special	5458,62	5821,11	7295,76	9727,4	7587,42	9433,39	7713,86	7949,99
8	9	tomato round	45061,71	33964,99	40330,61	68065	35763,36	52405,42	48945,04	66471,74
9	10	tomato special	1165,56	778,73	917,59	1092,95	983,37	991,39	948,5	1015,57
10	11	raspberry	3541,23	4768,77	5865,47	9469,58	8258,84	8567,7	8153,2	10924,75
11	12	tomato big	2092,26	2576,32	666,26	369,7	3136,34	1486,75	945,59	0
12	13	tomato plum	23723,91	22749,05	25832,75	18797,13	19206,82	20195,97	22695,01	20184
13	14	watermelon	61647,98	45973,18	79374,35	45007,91	13186,04	43000,66	66137,99	60073,04
14	15	mixed soft fruits	715,44	697,42	776,07	619,24	521,59	646,59	687,02	666,07
15	16	paprika long	2976,06	9516,56	9184,64	8323,43	10876,93	9489,02	9793,23	9496,78
16	17	tomato vine plum	4041,55	8130,89	9767,71	11050,35	10328,57	11192,35	11367,23	13221,94
17	18	blackberry	650,23	818,89	1012,35	1258,48	1366,99	1445,95	1381,43	1521,23
18	19	galia	17757,62	61424,99	114164,67	58267,54	52490,31	83944,39	91017,35	80846,27
19	20	others	10687,61	3931,33	15066,27	17746,15	8587,17	16457,29	14956,44	19272,85

Tabela de previsão de vendas por produtos para 8 semanas (quantidades em Kg)



Gráficos de variação das vendas por produtos, acrescentados da previsão estimada

Etapa 1 - Processo ETL - Os dados recolhidos foram lidos, tratados e integrados, com uniformização de grandezas, eliminação de redundâncias ou duplicados, e tratamento de valores nulos ou elegíveis. Foram recebidos dados desde janeiro 2018 até início de maio de 2022, num total de 315.474 registos de vendas, 59.244 produtos, traduzindo-se depois do ETL em 29 produtos únicos e válidos, e 52 clientes válidos e com operações registadas.

Etapa 2 - Validação - Foi definido fazer uma análise detalhada para os produtos que percentualmente tivessem maior representatividade e cuja soma percentual representasse pelo menos 95% das vendas globais, em quantidades vendidas (Kg), mas também em número de ordens de compra recebidas. Resultaram 19 produtos significativos a tratar individualmente, sendo os restantes tratados como um conjunto.

Etapa 3 - Aplicação SARIMA - Foi utilizada uma função automática em Python, para determinação dos parâmetros adequados a cada modelo SARIMA, aplicado a cada serie temporal dos produtos definidos. Os modelos foram treinados e testados fornecendo resultados consistentes que validam o processo desenvolvido, pese embora os erros verificados rondem os 26%.

Etapa 4 - ETL Fatores Exógenos - Como fatores exógenos selecionamos as condições metrológicas compostas por 39 parâmetros diferentes. Numa primeira abordagem, intuitivamente, foram selecionados de forma fixa a Temperatura diária, a Duração da luz solar e a Humidade do ar. Simultaneamente e aplicado um algoritmo de redução de variáveis, RFE – Recursive Feature Elimination, tentamos pelo método RFE automático encontrar os 3 parâmetros principais que mais influenciam a venda de cada produto. Verificou-se que os 3 parâmetros fixos previamente selecionados não eram os dominantes, assim como se verificou, que os 3 principais parâmetros obtidos pelo processo RFE variavam de produto para produto.

Recorrendo a algoritmos de classificação (KNN1/ KNN2/ RFC/ LRN/ SVM) e avaliando a precisão dos resultados obtidos para os dois métodos, ainda que com uma ligeira diferença foi escolhido o processo automático pelo algoritmo RFE pois apresenta em média melhor precisão em todos os classificadores, com valor de 74% comparado com 73% do processo intuitivo e fixo inicialmente abordado.

Etapa 5 - Aplicação SARIMAX - Foi então utilizado o modelo SARIMAX como forma de estudar as previsões de venda dos produtos, considerando a influência das condições meteorológicas. Também neste caso, foi utilizada uma função automática em Python, na determinação dos parâmetros adequados a cada modelo SARIMAX. Os modelos foram treinados e testados fornecendo resultados consistentes que validam o processo desenvolvido, verificando-se uma melhoria ligeira de 26% para 19% nos erros verificados.

Etapa 6 - Previsões - Como última fase, e aplicando os modelos desenvolvidos, são geradas em formato tabela e gráfico as previsões estimadas de vendas para os todos os produtos nas 8 semanas subsequentes.

CONCLUSÃO

A solução desenvolvida para a “The Greenery” em contexto real, consiste numa aplicação em Python, capaz de prever vendas de vários produtos num horizonte temporal de 8 semanas recorrendo aos dados históricos de vendas ocorridas. O resultado desta aplicação é uma tabela de fácil interpretação e utilização, conforme foi solicitado.

O trabalho apresentado, e partindo de uma necessidade concreta do mercado, é o resultado de um processo iterativo de aprendizagem e aplicação de conhecimento, traduzido numa solução possível de fácil aplicação na “The Greenery”.

BIBLIOGRAFIA & SITES CONSULTADOS

1) https://www.datascience-pm.com/crisp-dm - Informação relativa à metodologia CRISP-DM apresentação e implementação;
2) https://www.thegreenery.com/en - Informação geral e operacional da empresa em estudo a “The Greenery”;
3) https://stackoverflow.com - Exemplos e rotinas em python utilizadas no desenvolvimento do processo e programa realizado;
4) https://scikit-learn.org/stable/user_guide.html - Exemplos e algoritmos de Machine Learning utilizados no desenvolvimento do processo e programa realizado;
5) https://www.statsmodels.org/dev/examples - Exemplos e aplicações de modelos de previsão aplicados a séries temporais: