Compitiendo en Kaggle: Predicción de ventas

GRUPO DE R DE SEVILLA
Sevilla, 4 de Diciembre de 2018

Javier Tejedor Aguilera

javier.tejedor.aguilera@gmail.com

javier-tejedor-aguilera

kaggle ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Datos

kaggle

ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Data Description

- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what you are predicting)
- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval	DayOfWeek	Date	Sales	Open	Promo	StateHoliday	SchoolHoliday
1	c	a	1270	9	2008	0	NA	NA		2	01/01/2013	0	0	0	a	1
1	c	a	1270	9	2008	0	NA	NA		3	02/01/2013	5530	1	0	0	1
1	c	a	1270	9	2008	0	NA	NA		4	03/01/2013	4327	1	0	0	1
1	c	a	1270	9	2008	0	NA	NA		5	04/01/2013	4486	1	0	0	1
1	c	a	1270	9	2008	0	NA	NA		6	05/01/2013	4997	1	0	0	1
1115	d	c	5350	NA	NA	1	22	2012	Mar,Jun,Sept,Dec	7	13/09/2015	NA	0	0	0	0
1115	d	c	5350	NA	NA	1	22	2012	Mar,Jun,Sept,Dec	1	14/09/2015	NA	1	1	0	0
1115	d	c	5350	NA	NA	1	22	2012	Mar,Jun,Sept,Dec	2	15/09/2015	NA	1	1	0	0
1115	d	c	5350	NA	NA	1	22	2012	Mar,Jun,Sept,Dec	3	16/09/2015	NA	1	1	0	0
1115	d	c	5350	NA	NA	1	22	2012	Mar,Jun,Sept,Dec	4	17/09/2015	NA	1	1	0	0

Métrica de evaluación

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Overview

Description

Evaluation

Prizes

Timeline

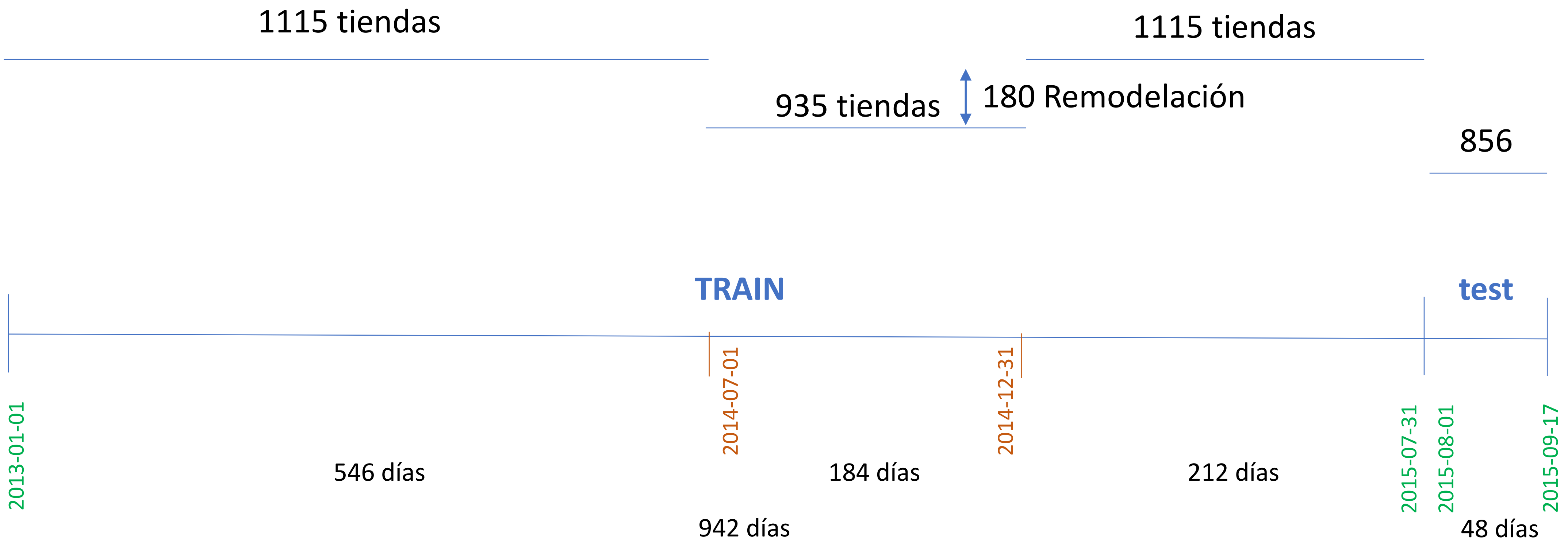
Submissions are evaluated on the Root Mean Square Percentage Error (RMSPE). The RMSPE is calculated as

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2},$$

where y_i denotes the sales of a single store on a single day and yhat_i denotes the corresponding prediction. Any day and store with 0 sales is ignored in scoring.

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data



Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Estrategia de validación / crossvalidación

kaggle
ROSSMANN

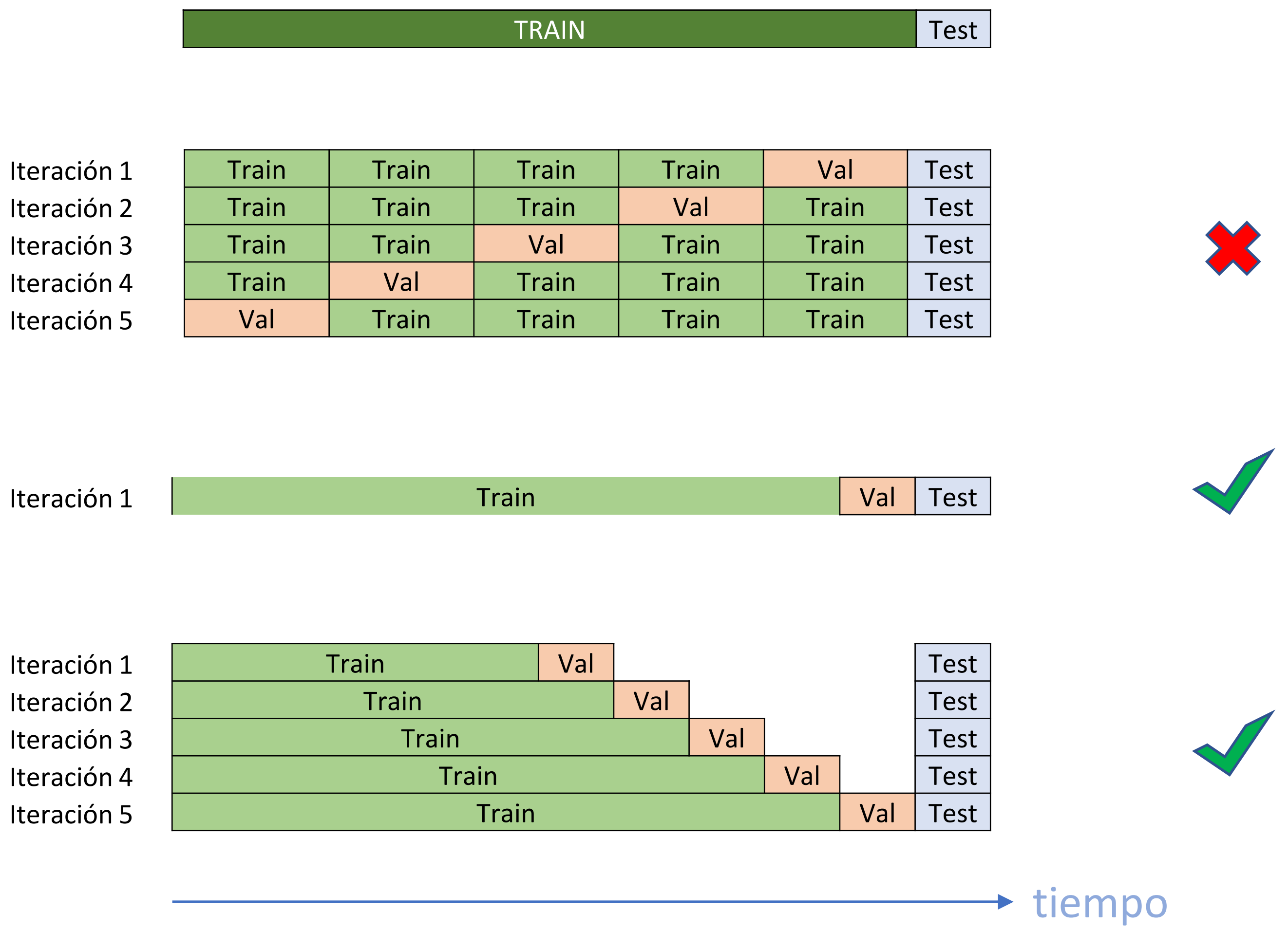
Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Crossvalidación 5-Fold Random

Validación temporal

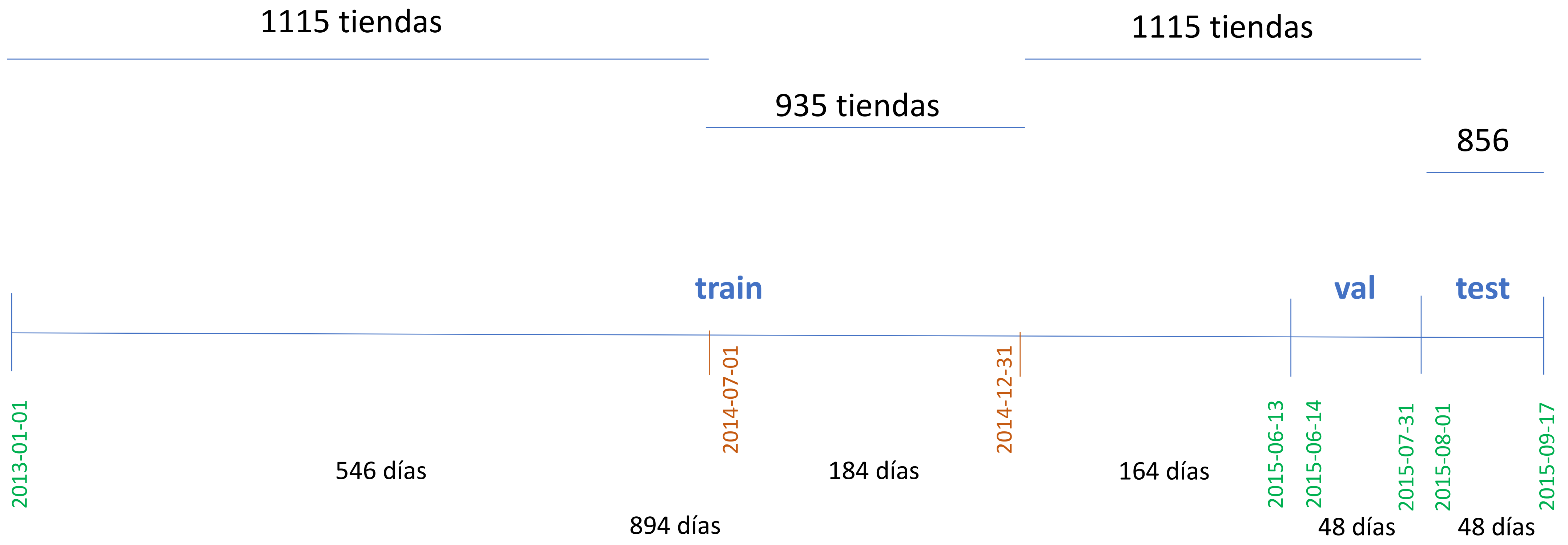
Crossvalidación 5-Fold temporal



El dataset de validación debe mimetizar el test

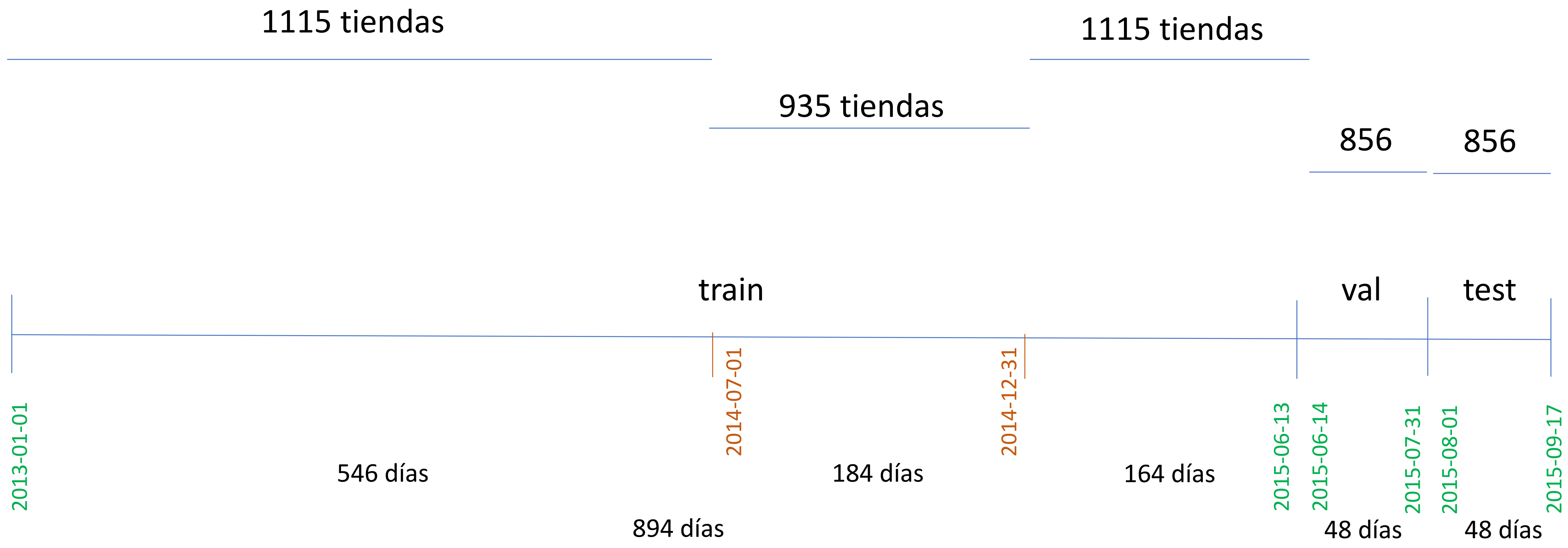
Rossmann Store Sales

Forecast sales using store, promotion, and competitor data



Rossmann Store Sales

Forecast sales using store, promotion, and competitor data



kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Ingeniería de Variables
(Modelo Gradient Boosting Decision Trees)

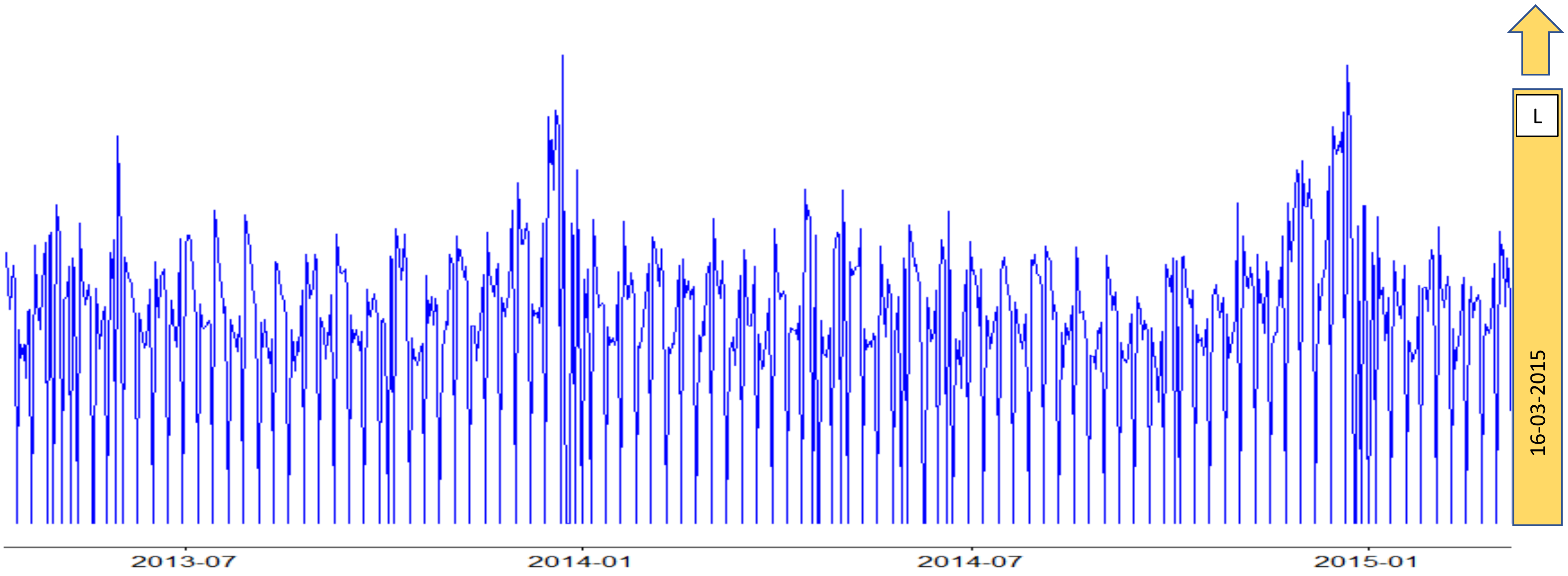
kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

dia	16
DayOfWeek	L
sem	12
mes	3
anio	2015

Tienda: T
Fecha: 16-03-2015
Variables: dia, sem, mes, anio

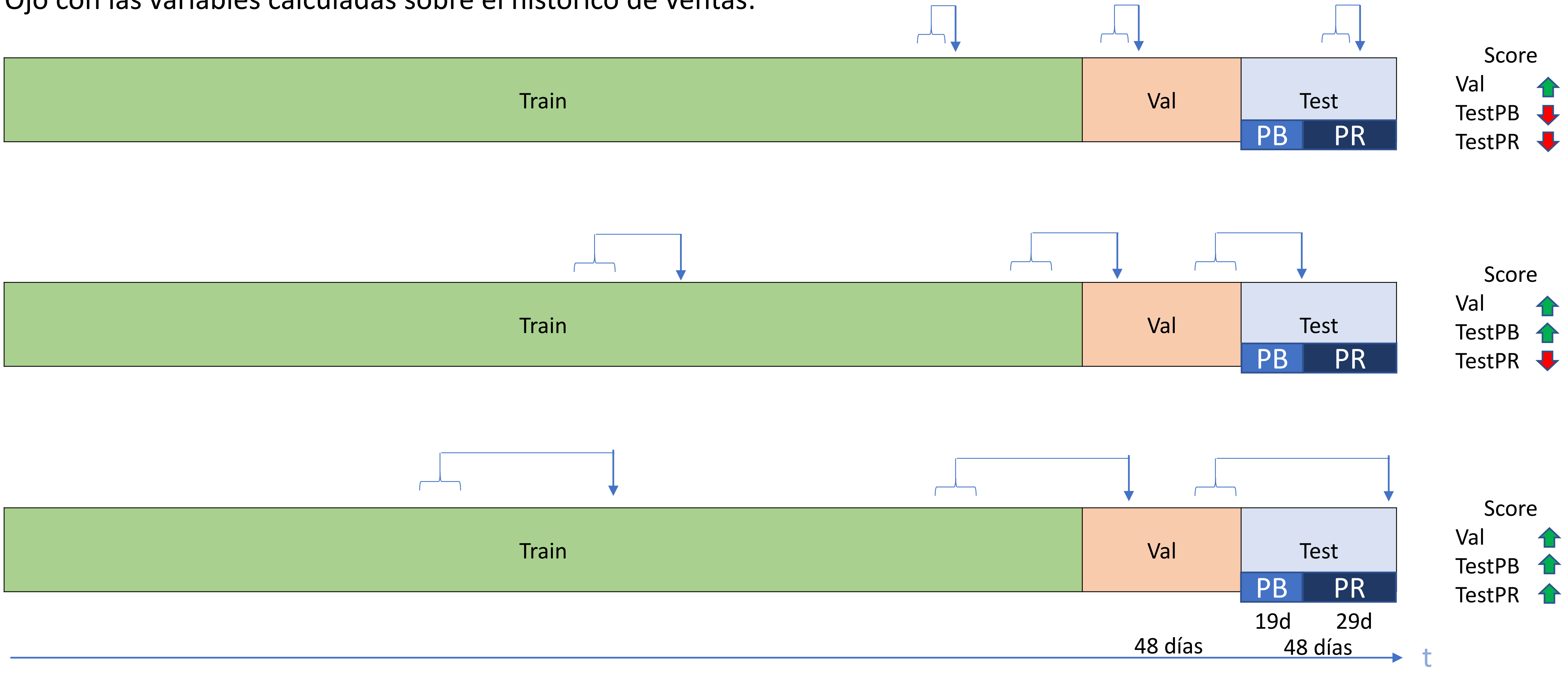


kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Ojo con las variables calculadas sobre el histórico de ventas:



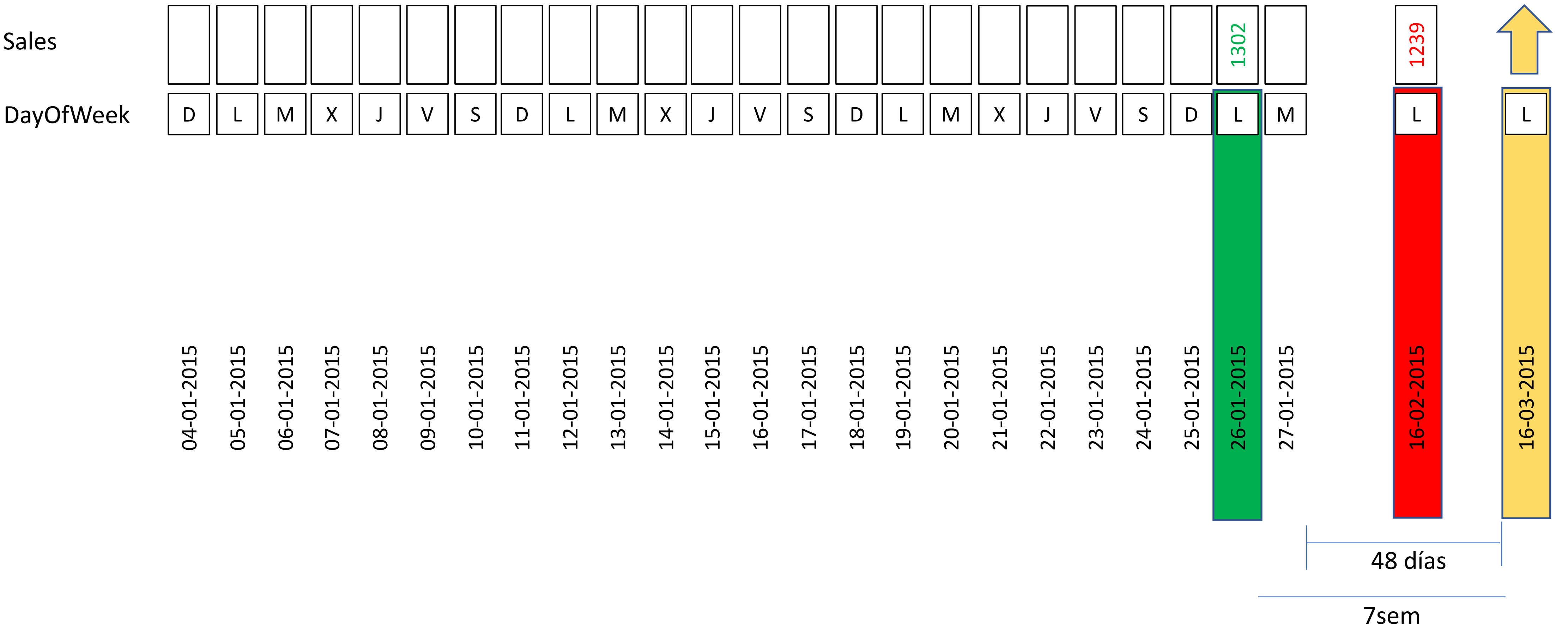
kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Tienda: T
Fecha: 16-03-2015
Variables: Sales_DOW_lag_7w

Sales_DOW_lag_7w **1032**

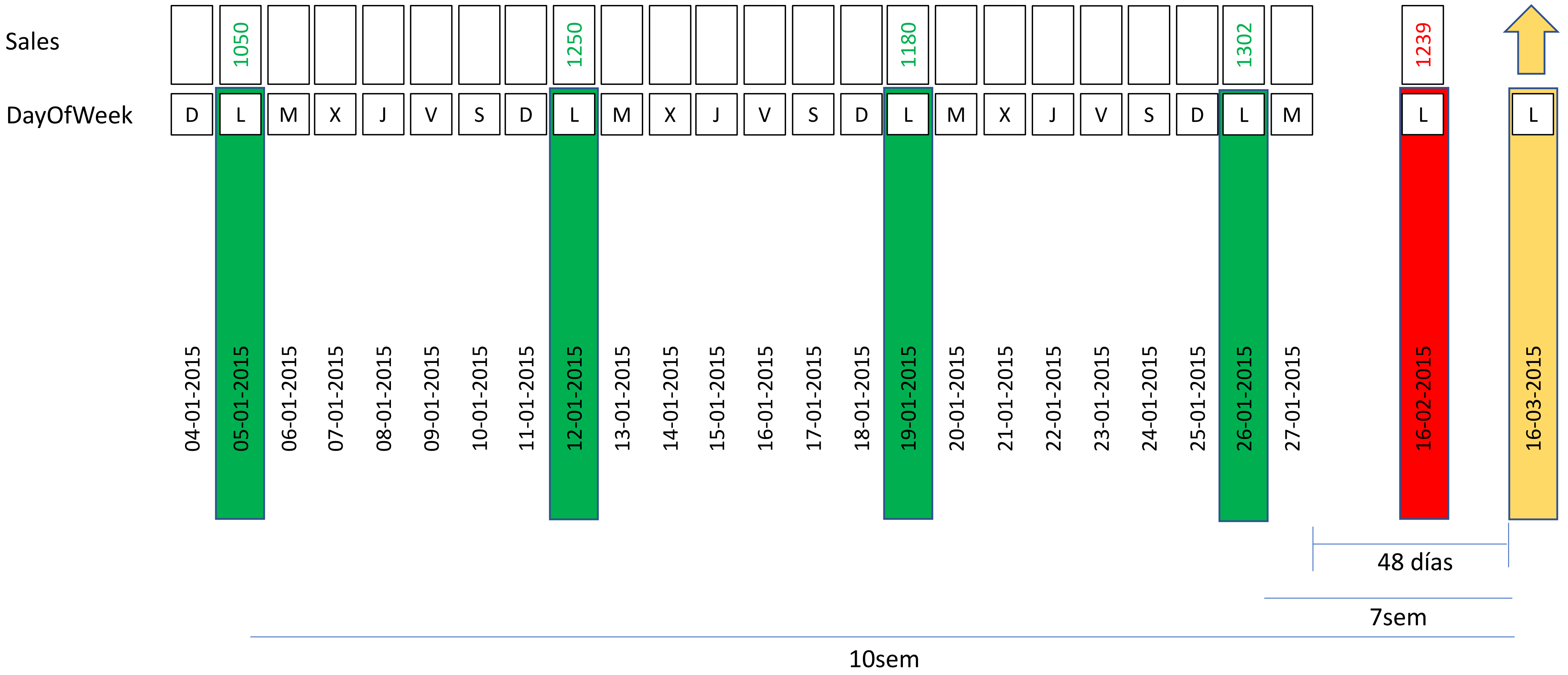


Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Media(1050,1250,1180,1302)	1195.5
Mediana(1050,1250,1180,1302)	1215
Min(1050,1250,1180,1302)	1050
Max(1050,1250,1180,1302)	1302
Sd(1050,1250,1180,1302)	109.12

Tienda: T
Fecha: 16-03-2015
Variables: Sales_mean_4_DOW, Sales_min_4_DOW, Sales_max_4_DOW, Sales_sd_4_DOW



kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Media(1250,1180,1302)	**1244**
Mediana(1250,1180,1302)	**1250**
Min(1250,1180,1302)	**1180**
Max(1250,1180,1302)	**1302**
Sd(1250,1180,1302)	**61,22**

Tienda: T
Fecha: 16-03-2015
Variables: Mean_Sales_Grupo_lag,

	04-01-2015	05-01-2015	06-01-2015	07-01-2015	08-01-2015	09-01-2015	10-01-2015	11-01-2015	12-01-2015	13-01-2015	14-01-2015	15-01-2015	16-01-2015	17-01-2015	18-01-2015	19-01-2015	20-01-2015	21-01-2015	22-01-2015	23-01-2015	24-01-2015	25-01-2015	26-01-2015	27-01-2015	…	16-02-2015	…	16-03-2015
Sales		1050							1250							1180							1302			1239		
DayOfWeek	D	L	M	X	J	V	S	D	L	M	X	J	V	S	D	L	M	X	J	V	S	D	L	M		L		L
StateHoliday	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N		N		N
SchoolHoliday	S	S	S	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	•••	N	•••	N
Open	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		S		S
Promo	N	S	S	S	S	S	S	S	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N		N		N

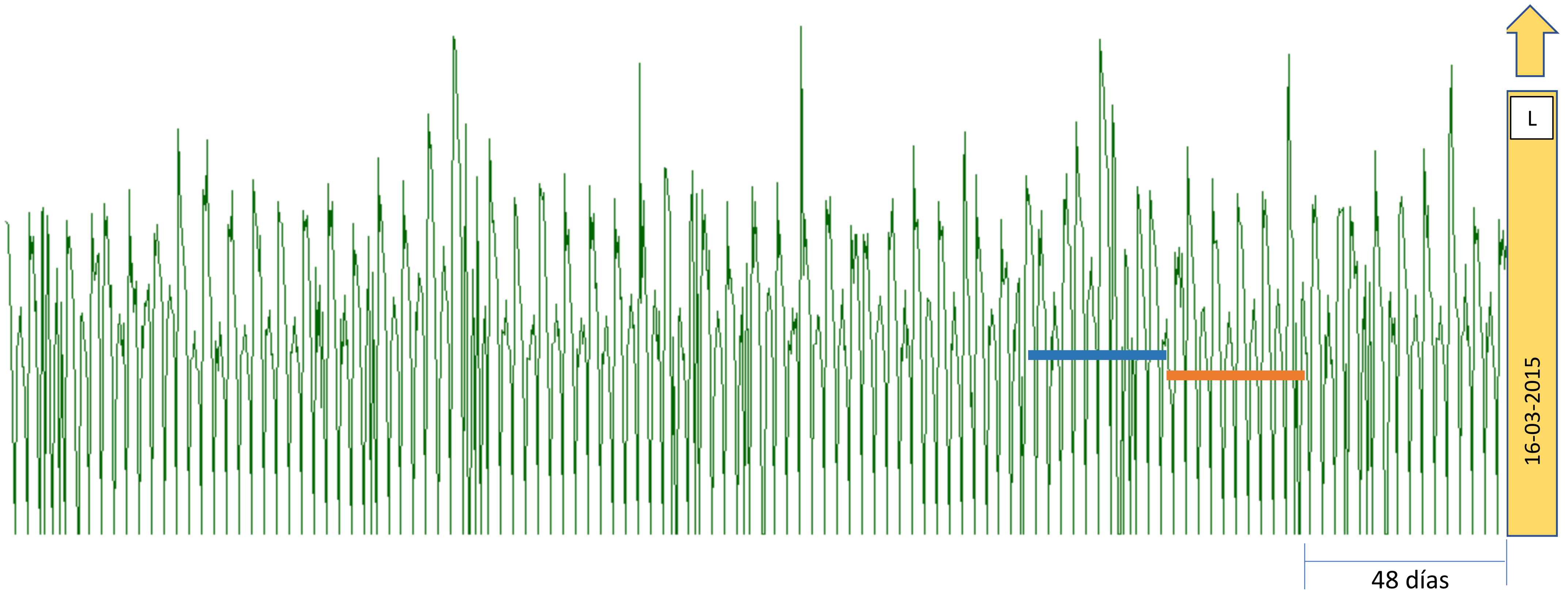
48 días

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Tienda: T
Fecha: 16-03-2015
Variables: Sales_mean_1m, Sales_mean_2m, ratio_mean_1m_2m

Sales_mean_1m	510
Sales_mean_2m	520
ratio_mean_1m_2m	0,98



kaggle

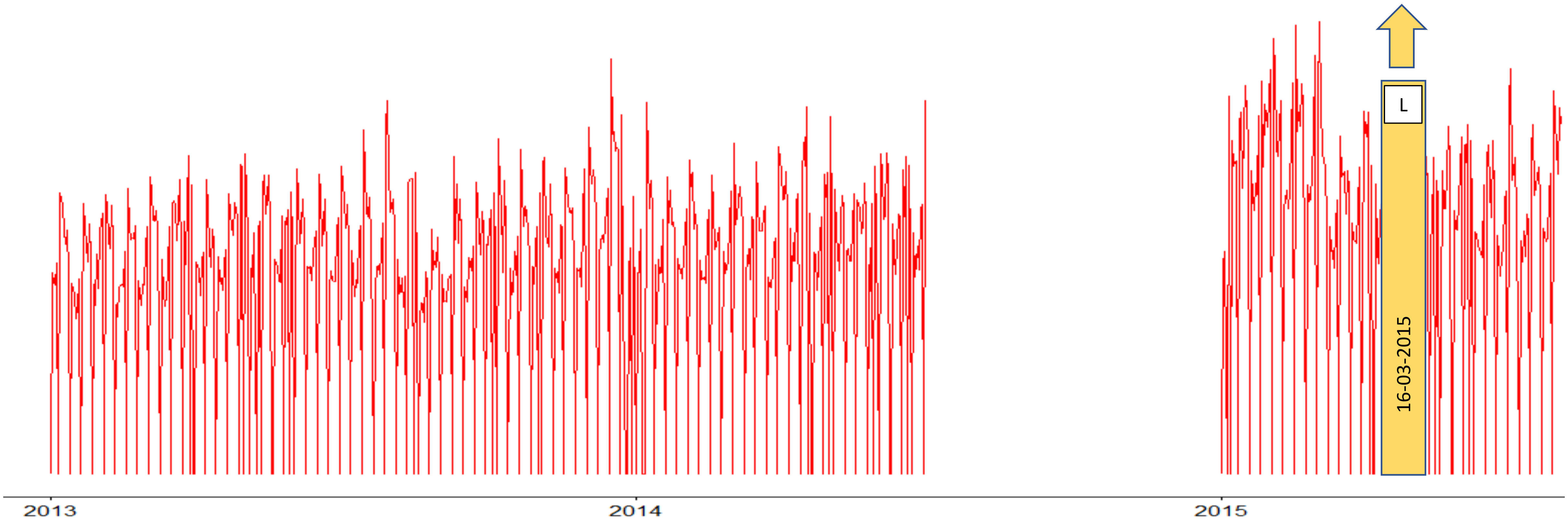
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Tienda: T
Fecha: 16-03-2015
Variables: tienda_con_remodelación, DiasDesdeRemodelacion, DiasParaRemodelacion

tienda_con_remodelación	1
DiasDesdeRemodelación	75
DiasParaRemodelación	3395

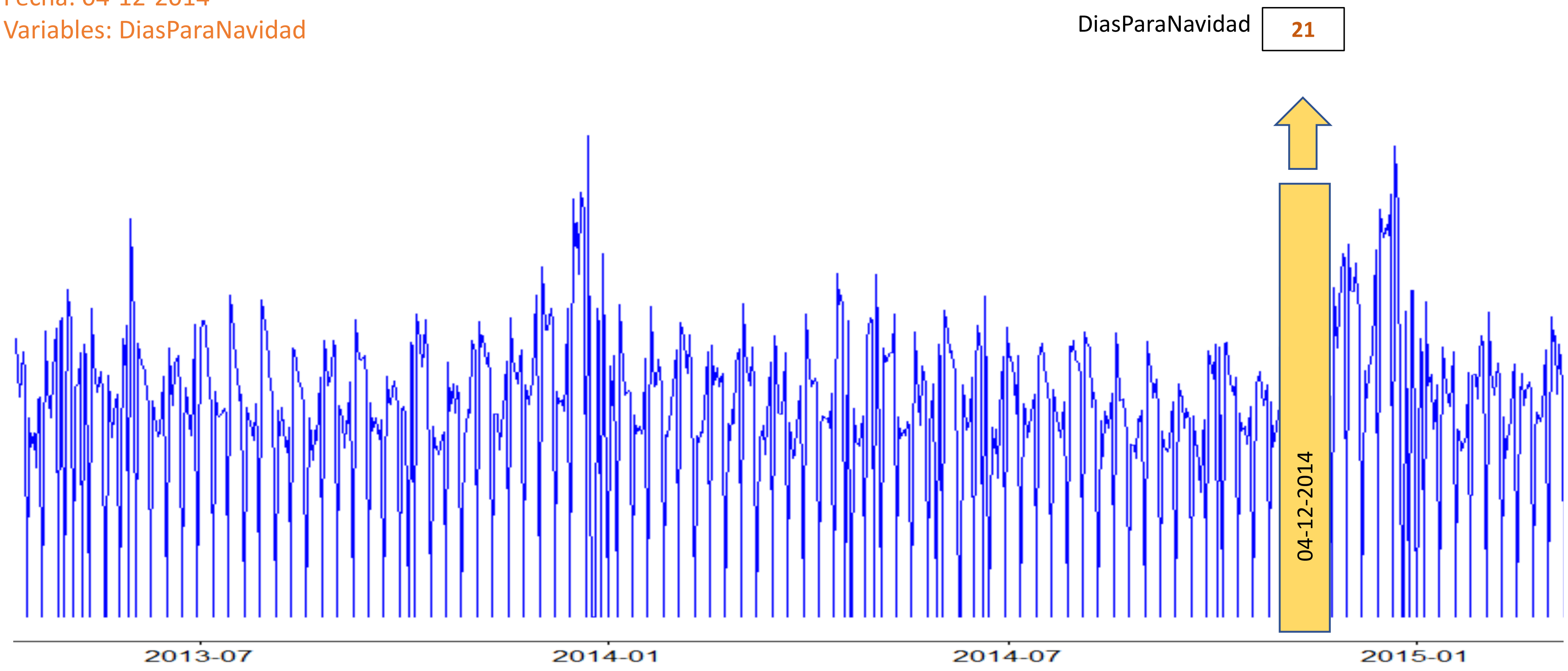


kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Tienda: T
Fecha: 04-12-2014
Variables: DiasParaNavidad



kaggle

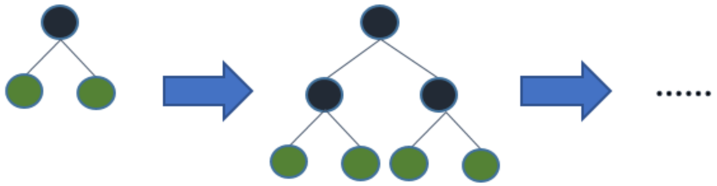
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Modelo

dmlc
XGBoost



Level-wise tree growth

GBDT Hyper Parameter Tuning

Hyper Parameter	Tuning Approach	Range	Note
# of Trees	Fixed value	100-1000	Depending on datasize
Learning Rate	Fixed => Fine Tune	[2 - 10] / # of Trees	Depending on # trees
Row Sampling	Grid Search	[.5, .75, 1.0]	
Column Sampling	Grid Search	[.4, .6, .8, 1.0]	
Min Leaf Weight	Fixed => Fine Tune	3/(% of rare events)	Rule of thumb
Max Tree Depth	Grid Search	[4, 6, 8, 10]	
Min Split Gain	Fixed	0	Keep it 0

Best GBDT implementation today: https://github.com/tqchen/xgboost
by **Tianqi Chen** (U of Washington)

DataRobot

Xgboost	• Eta • Gamma • Max_depth • Min_child_weight • Subsample • Colsample_bytree • Lambda • alpha	• 0.01,0.015, 0.025, 0.05, 0.1 • 0.05-0.1,0.3,0.5,0.7,0.9,1.0 • 3, 5, 7, 9, 12, 15, 17, 25 • 1, 3, 5, 7 • 0.6, 0.7, 0.8, 0.9, 1.0 • 0.6, 0.7, 0.8, 0.9, 1.0 • 0.01-0.1, 1.0 , RS* • 0, 0.1, 0.5, 1.0 RS*

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

Ideas de mejora

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

1. Generar N modelos, uno por día a predecir

Idea propuesta por **Danijel Kivaranovic** (Kaggle Grandmaster) en la competición Recruit Restaurant Visitor Forecasting de Kaggle y que expuso en Kaggle Days Varsovia.

Recruit Restaurant Visitor Forecasting
Predecir los visitantes en 314 restaurantes en los próximas 39 días

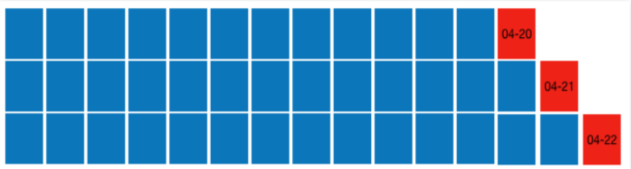
kaggle
ROSSMANN

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

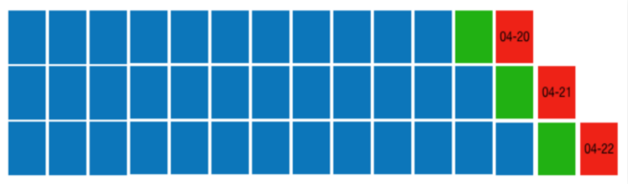
Modelo 1

Data that can be used is blue. Data that must be ignored is gray. Data that is not available is black.



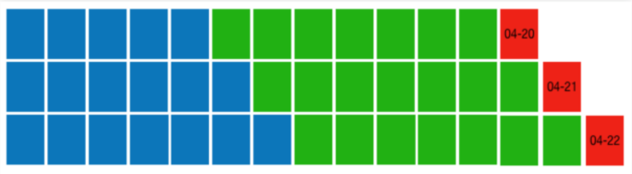
Data used to compute the feature in green.

Day feature



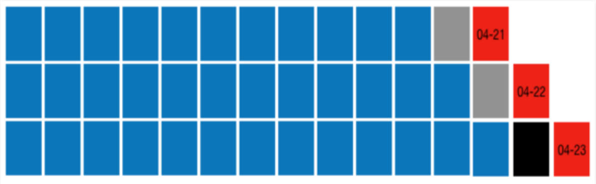
Data used to compute the feature in green.

Week feature



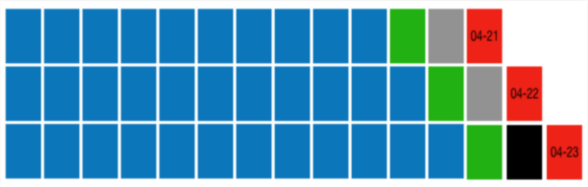
Modelo 2

Data that can be used is blue. Data that must be ignored is gray. Data that is not available is black.



Data used to compute the feature in green.

Day feature

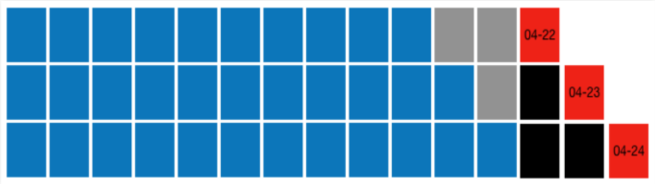


Data used to compute the feature in green.

Week feature

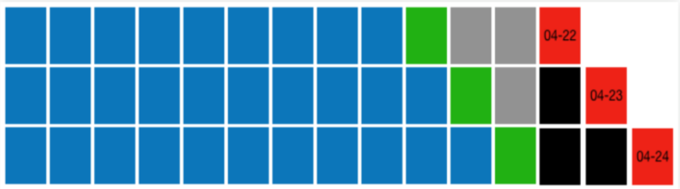
Modelo 3

Data that can be used is blue. Data that must be ignored is gray. Data that is not available is black.



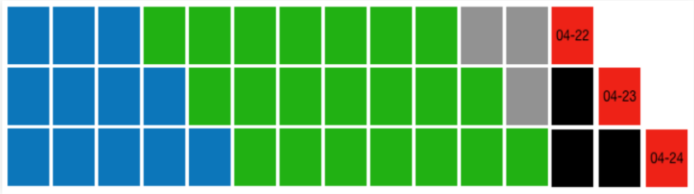
Data used to compute the feature in green.

Day feature



Data used to compute the feature in green.

Week feature

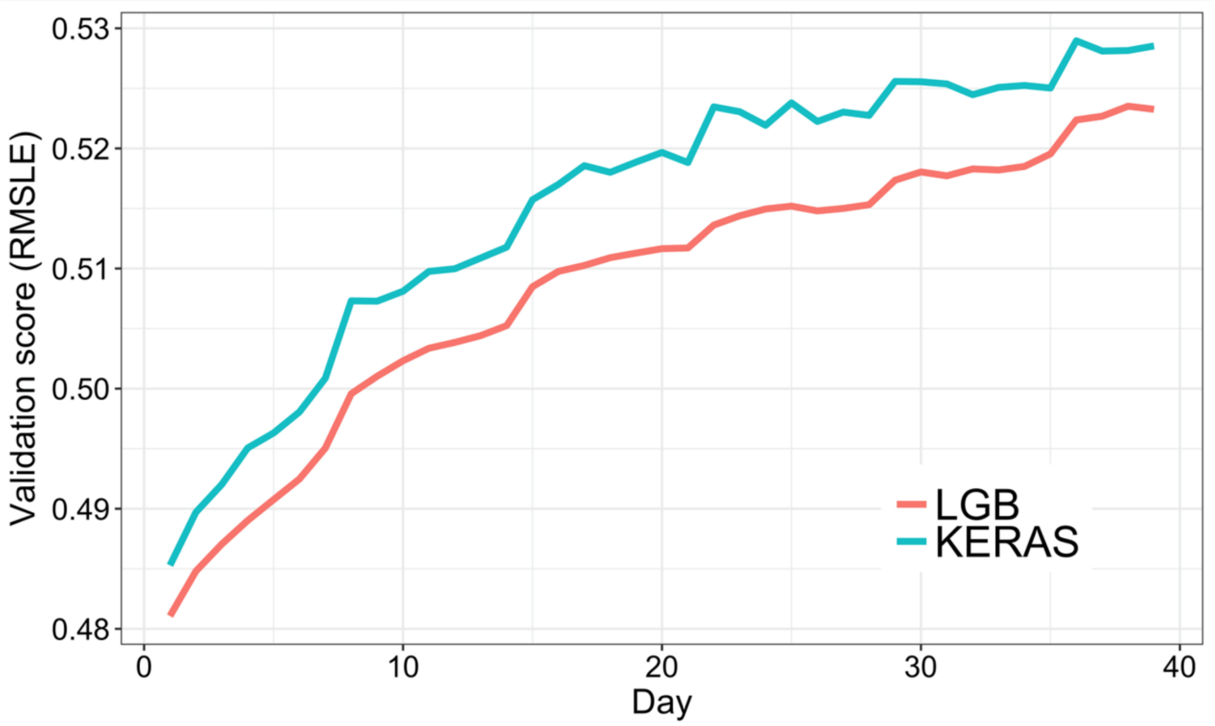


Nota: Imágenes extraídas de las slides de la presentación de **Danijel Kivaranovic** en Kaggle Days Warsaw

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

39 días -> 39 Modelos -> 39 datasets de validación -> 39 scores de validación



Nota: Imagen extraída de la presentación de **Danijel Kivaranovic** en Kaggle Days Warsaw

Como era de esperar, cuanto más lejana en el tiempo la predicción, mayor es el error.

kaggle ROSSMANN **Rossmann Store Sales**

Forecast sales using store, promotion, and competitor data

Más ideas de mejora

1. Mejorar ingeniería de variables
    - Mejorar algunas de las variables calculadas
    - Cambiar codificación de las categóricas: oneHotEncoding, LabelEncoding, Frecuencia, etc.
    - Nuevas variables

2. Bagging
    - Promediar modificando algunos hiperparámetros
    - Promediar modificando la forma de cálculo de algunas variables (cambiando codificación categóricas, etc.)

3. Stacking
    - Combinar las predicciones de varios modelos (RF, RNN, LR, ARIMA, …) sacando lo mejor de cada uno. Cuanto mayor diversidad, mejor.

Muchas gracias
por vuestra atención!