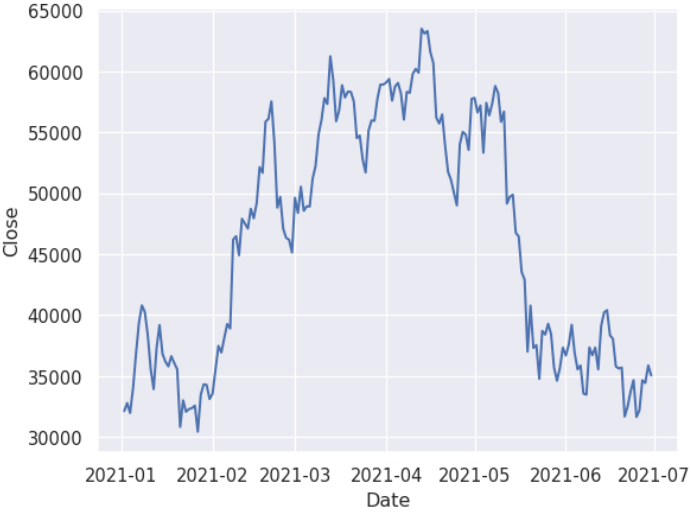
Regresión polinomial

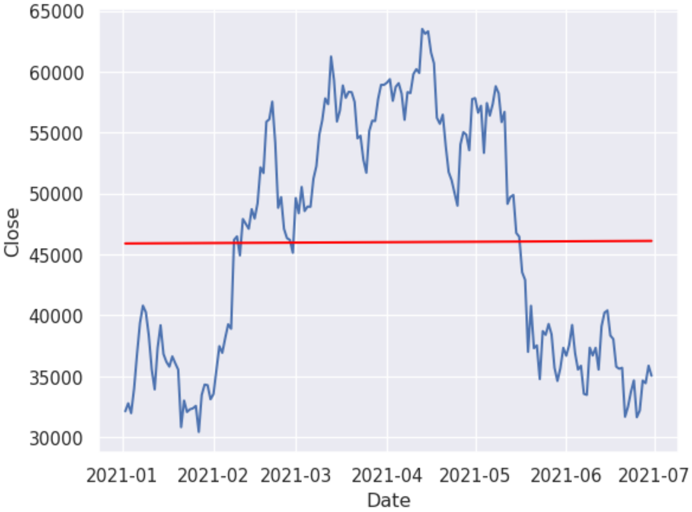
Laboratorio de Datos, IC - FCEN - UBA - 1er. Cuatrimestre 2025

Regresión Polinomial

Teníamos los datos del precio de Bitcoin durante cierto periodo



El ajuste lineal no explica muy bien la evolución del precio...



$$Y = 45900 + 1{,}18X$$
$$R^2 \approx 3{,}81 \times 10^{-5}$$

Obs: el promedio del precio en este periodo fue de U$D46005

Cuando la recta no ajusta bien a los datos, podemos intentar ajustarlos con un **polinomio** de grado más grande.

Recordemos: polinomio de grado n

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_n X^n$$

Supongamos que queremos aproximar los datos con un polinomio de grado *a lo sumo* 5:

$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$

Supongamos que queremos aproximar los datos con un polinomio de grado *a lo sumo* 5:

$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$

Tal cual hicimos con regresión lineal, queremos encontrar los valores de $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ que minimicen los residuos:

$$y_i - P(x_i)$$

Para minimizar los residuos, podemos usar **Cuadrados Mínimos**. Es decir, igual que hicimos con Regresión Lineal, queremos encontrar los β que minimicen la suma de los cuadrados de los residuos:

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - P(x_i))^2$$

Para minimizar los residuos, podemos usar **Cuadrados Mínimos**. Es decir, igual que hicimos con Regresión Lineal, queremos encontrar los β que minimicen la suma de los cuadrados de los residuos:

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - P(x_i))^2$$

Las medidas de desempeño del modelo son análogas a las de Regresión Lineal:

$$R^2 = \frac{\sum_{i=1}^{n} (P(x_i) - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad ECM = \frac{1}{n} \sum_{i=1}^{n} (y_i - P(x_i))^2$$

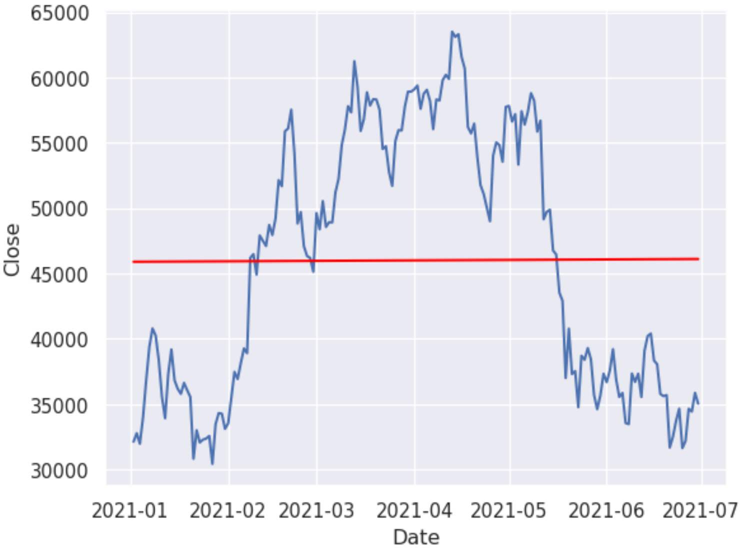
Si todo viene siendo muy parecido a Regresión Lineal, entonces deben haber fórmulas *sencillas* para calcular cada β…

Si todo viene siendo muy parecido a Regresión Lineal, entonces deben haber fórmulas *sencillas* para calcular cada β...

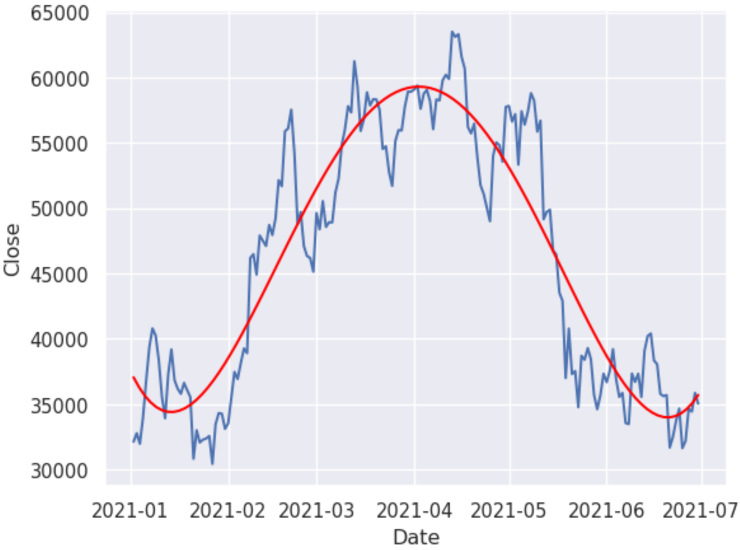
Si bien al buscar los mínimos de $RSS(\beta)$ también se obtiene un sistema de ecuaciones lineales, para calcular el valor de los β necesitamos herramientas que se ven en Álgebra Lineal Computacional.

Sin embargo, gracias a `seaborn`, la visualización es sencilla:

```
.add(so.Line(color='red'), so.PolyFit(1))
```



```
.add(so.Line(color='red'), so.PolyFit(5))
```



¿Cómo elegimos el grado del polinomio?

El grado del polinomio es el parámetro del modelo. No hay una fórmula o una regla que nos diga cuál es el grado apropiado para nuestros datos.

¿Cómo elegimos el grado del polinomio?

El grado del polinomio es el parámetro del modelo. No hay una fórmula o una regla que nos diga cuál es el grado apropiado para nuestros datos.

Sin embargo, tenemos algunas opciones:

- dividir nuestros datos en un conjunto de entrenamiento y en uno de testeo para probar cuál es el grado que mejor se desempeña.
- tener algún conocimiento específico respecto a nuestros datos.

¿Cómo elegimos el grado del polinomio?

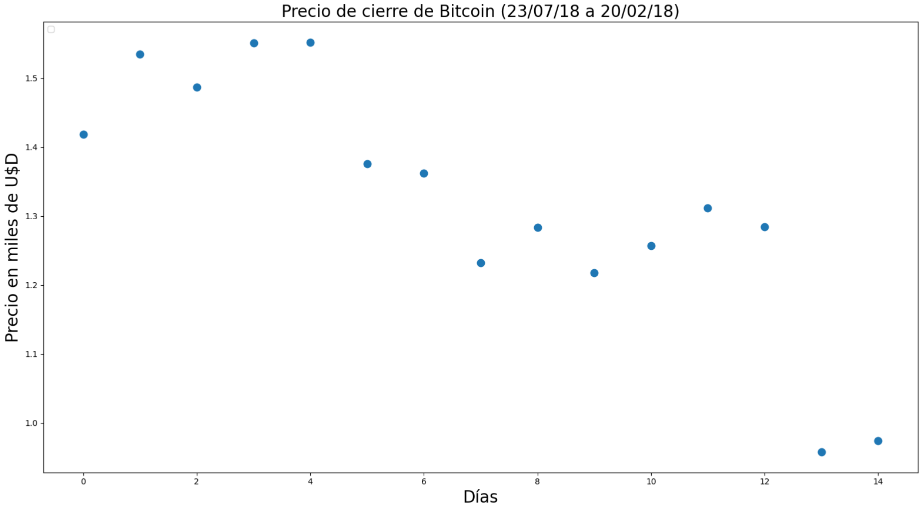
El grado del polinomio es el parámetro del modelo. No hay una fórmula o una regla que nos diga cuál es el grado apropiado para nuestros datos.

Sin embargo, tenemos algunas opciones:

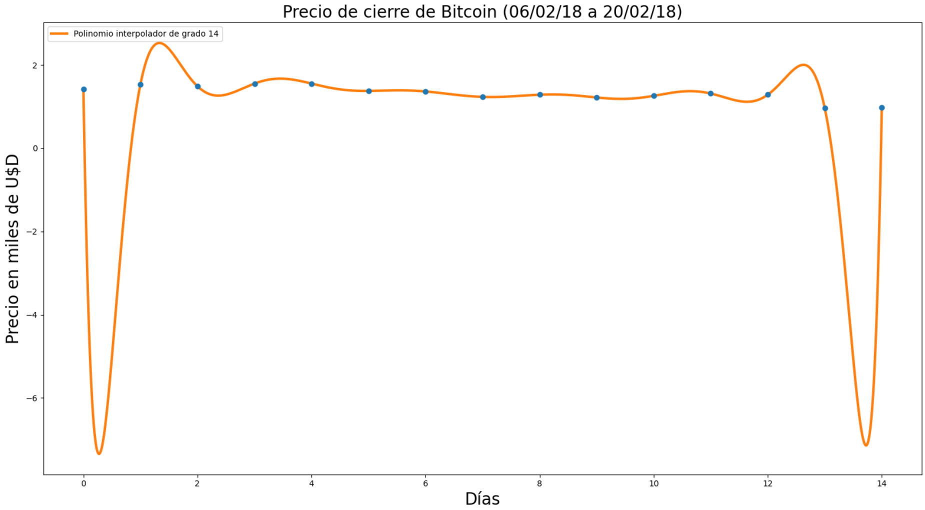
- dividir nuestros datos en un conjunto de entrenamiento y en uno de testeo para probar cuál es el grado que mejor se desempeña.
- tener algún conocimiento específico respecto a nuestros datos.

Obs: a mayor grado, más complejo es el modelo y mayores son los riesgos de *overfitting* y de tener un problema *mal condicionado*.

El riesgo del *overfitting*



El riesgo del *overfitting*



Bonus: Aplicación de Regresión Lineal

Media Movil de Cuadrados Minimos

Least Squares Moving Average (LSMA)

Media Movil de Cuadrados Minimos

Least Squares Moving Average (LSMA)

En el mercado bursátil, la LSMA es utilizada como un indicador:

- ”Suaviza” el movimiento del precio del activo por lo que captura mejor la **tendencia**. No tiene como objetivo predecir el precio.
- Temporalmente, se mueve un poco por detrás del precio del activo (*lag*).
- El cruce de dos LSMA de distintos periodos puede ser una señal de compra o de venta.

¿Cómo calculamos la LSMA?

Supongamos que tenemos nuestros datos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

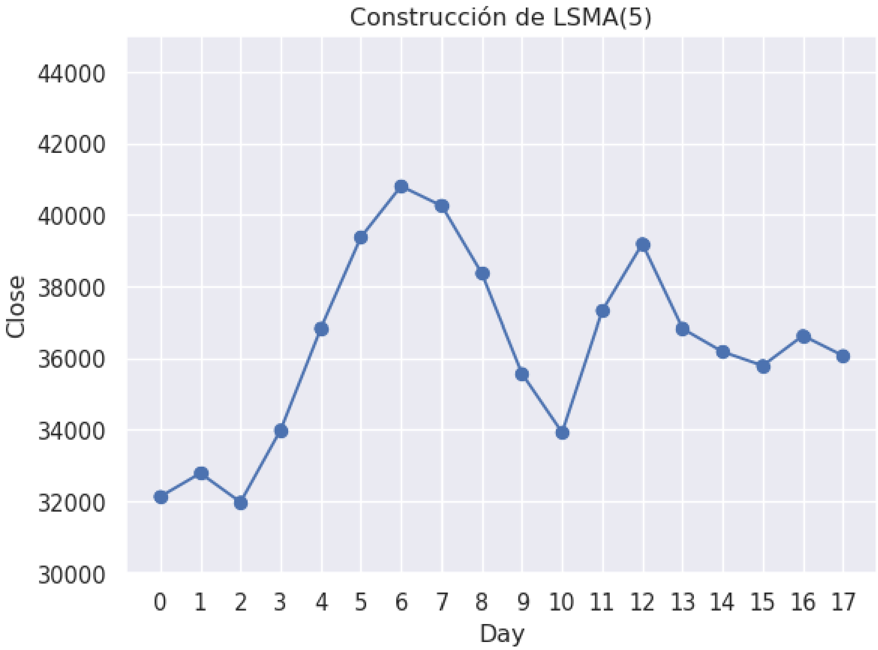
Lo primero que hay que hacer es fijar un periodo de tiempo k.

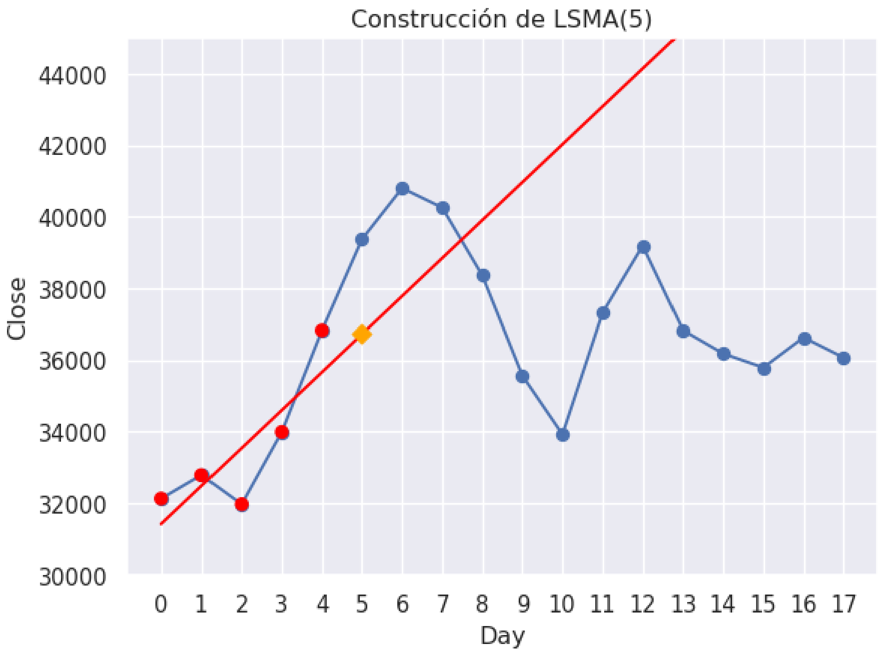
Notamos como $\hat{y}_i$ a la estimación de LSMA para x_i y se calcula de la siguiente manera:

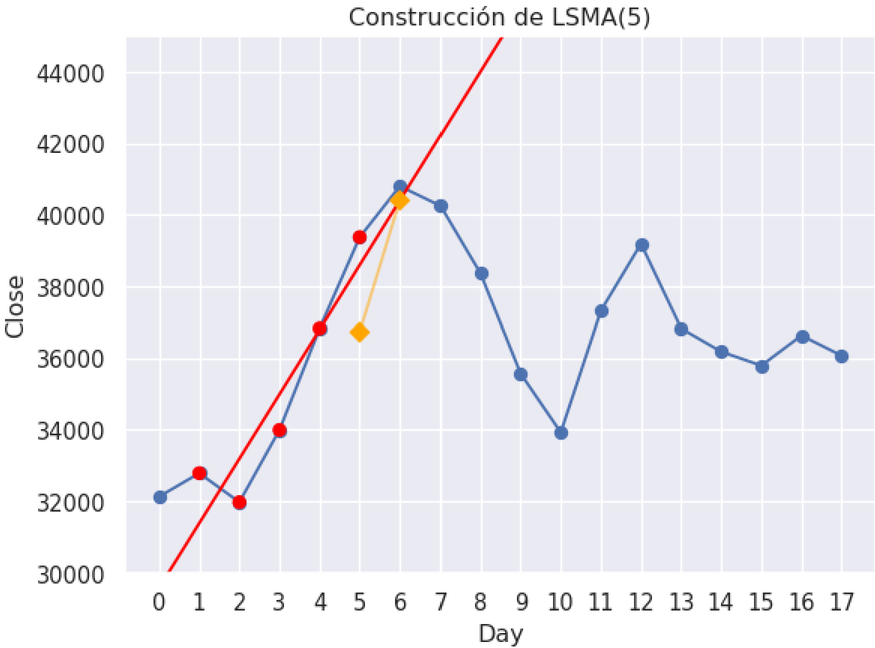
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

donde β_0 y β_1 son los coeficientes de la **Regresión Lineal sobre los anteriores k datos:**

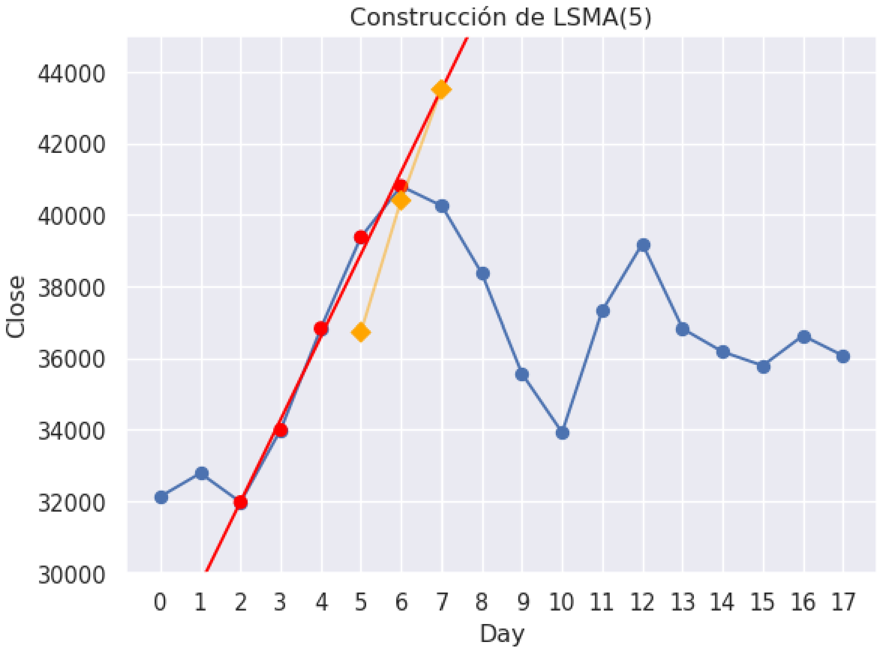
$$\{(x_{i-k-1}, y_{i-k-1}), (x_{i-k}, y_{i-k}), \ldots, (x_{i-1}, y_{i-1})\}$$

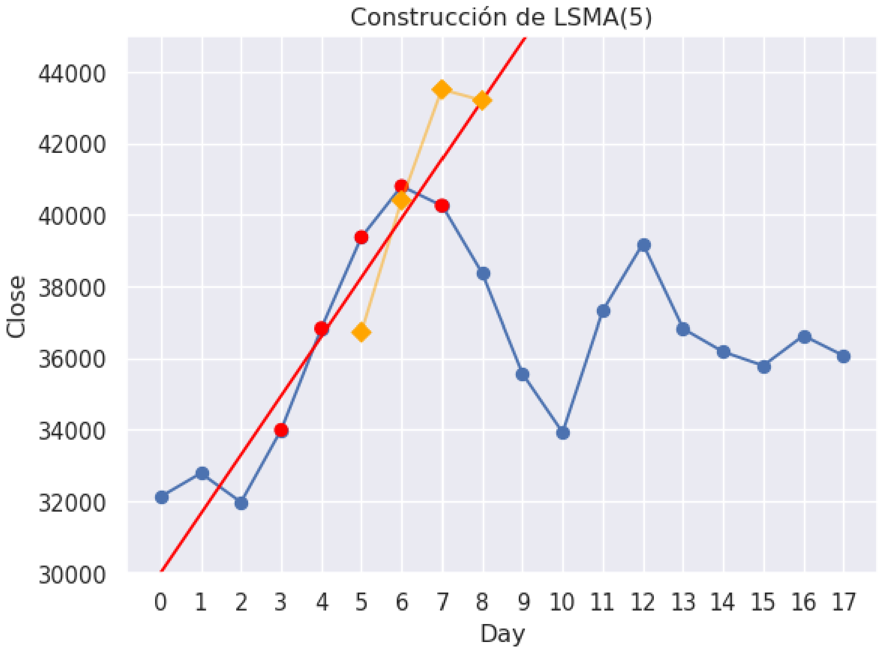

Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day

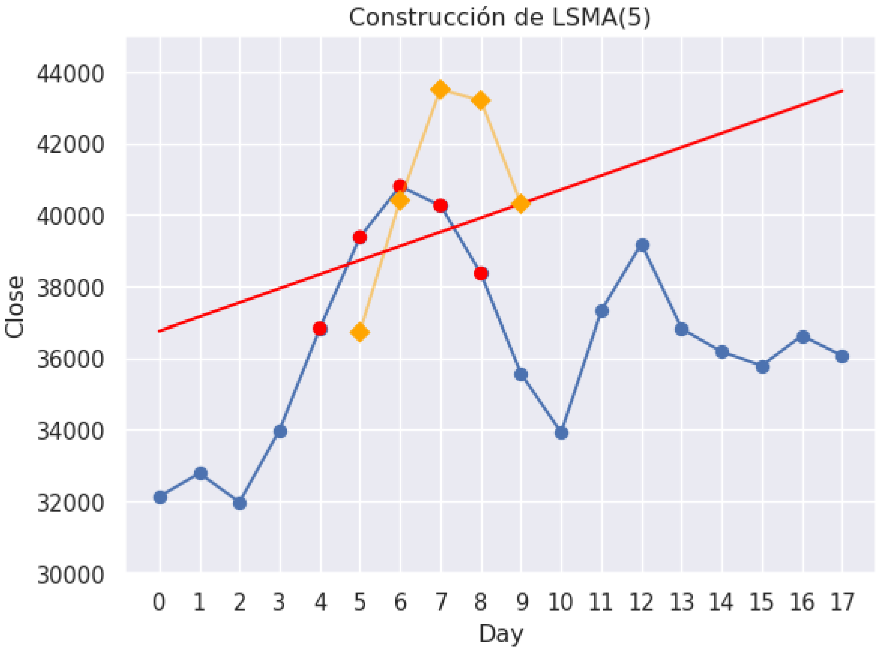

Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

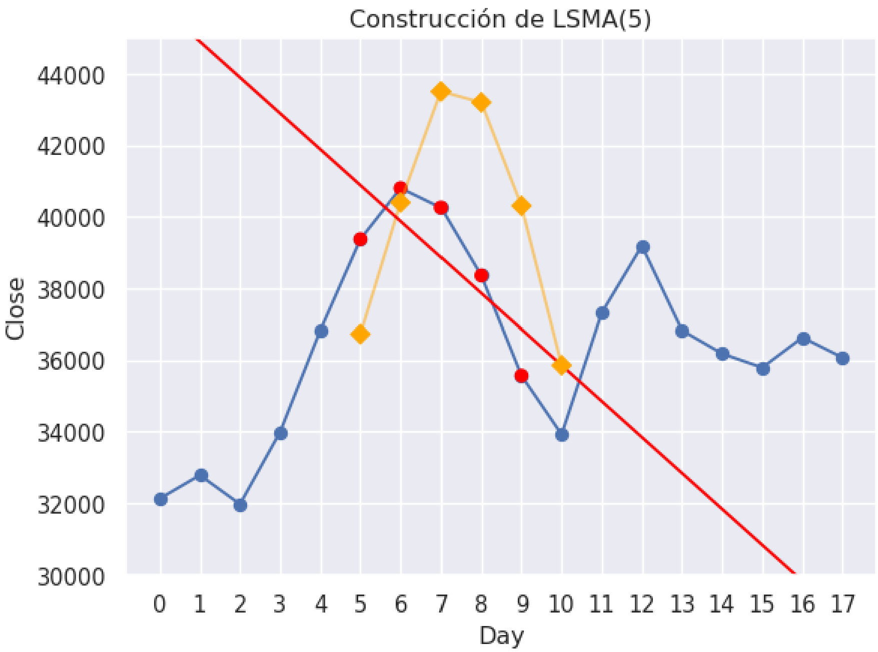


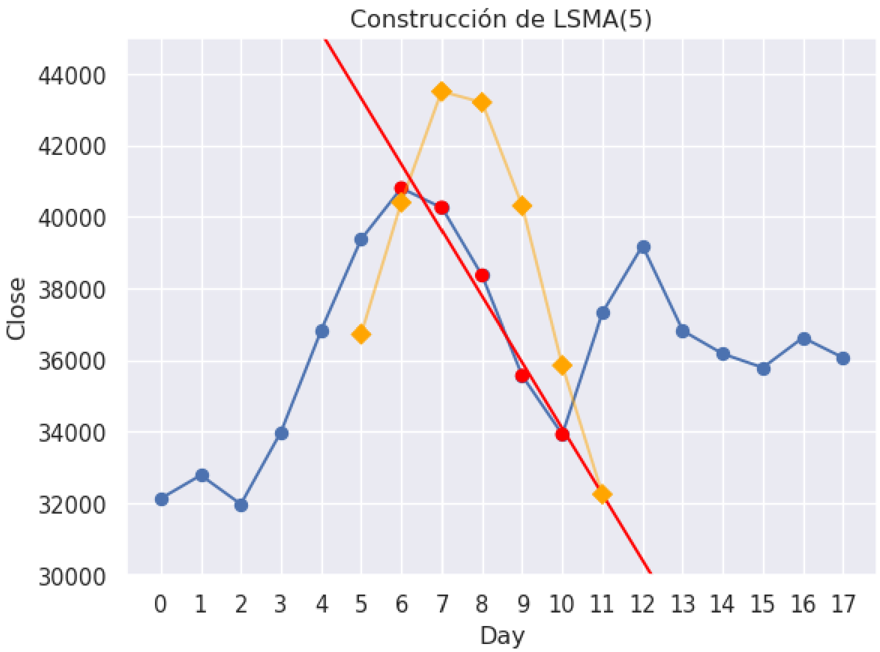
Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day

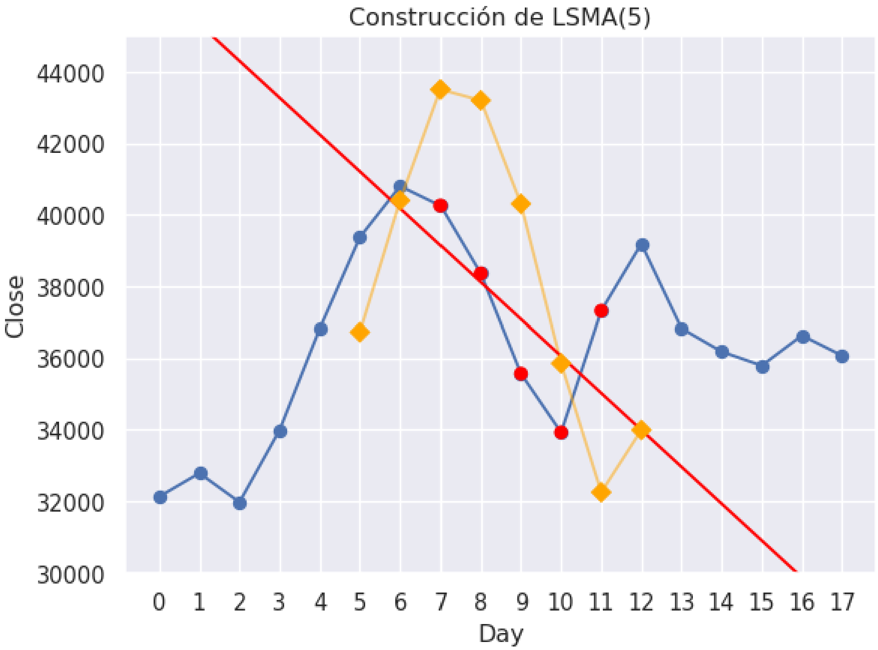

Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day

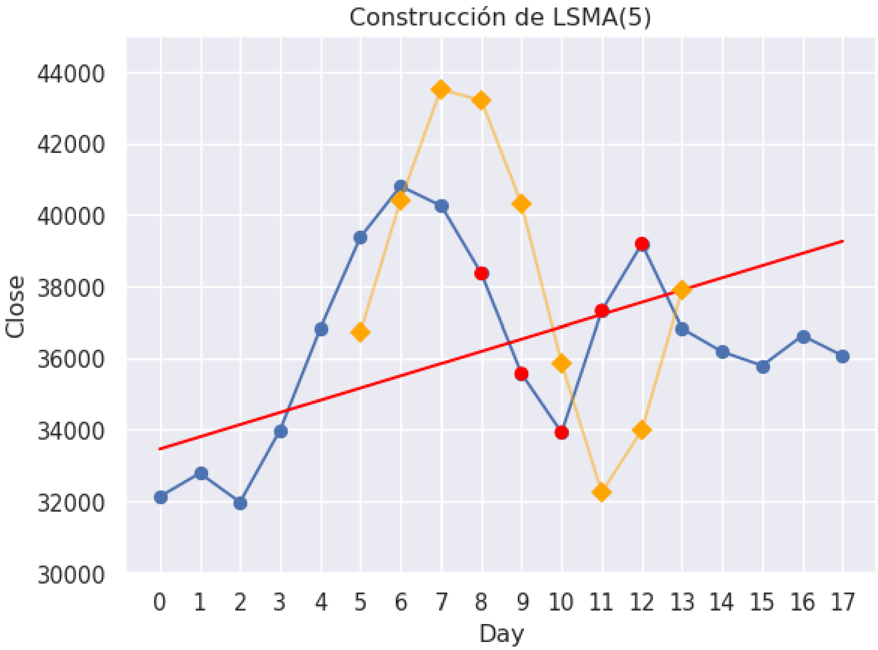

Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Day

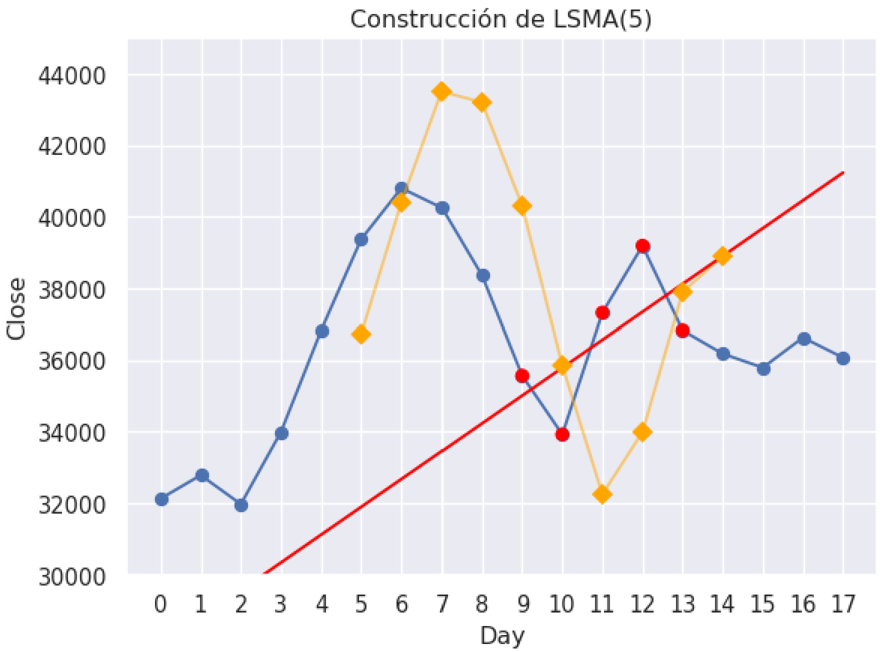

Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

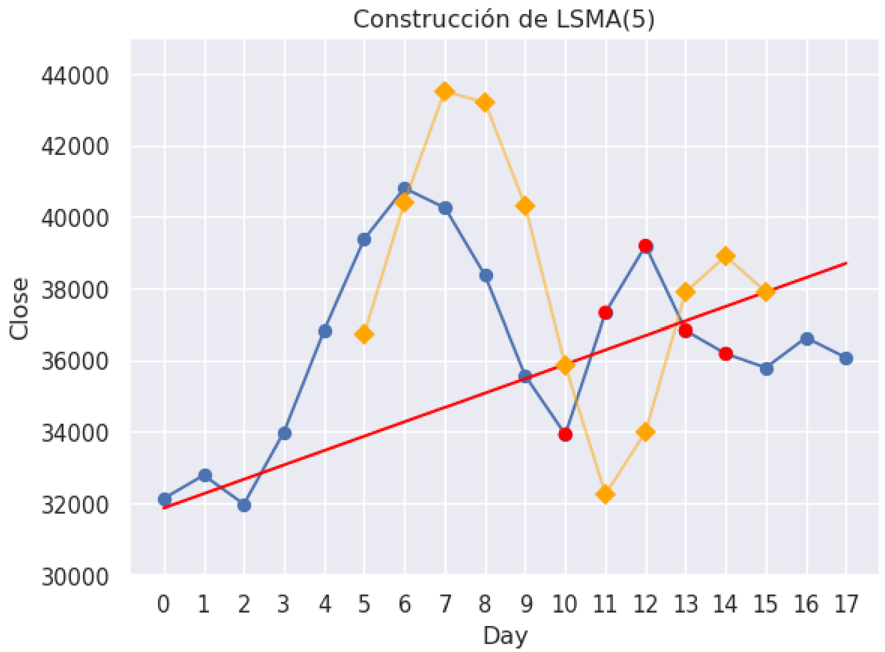

Construcción de LSMA(5)
44000
42000
40000
38000
36000
34000
32000
30000
Close
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day

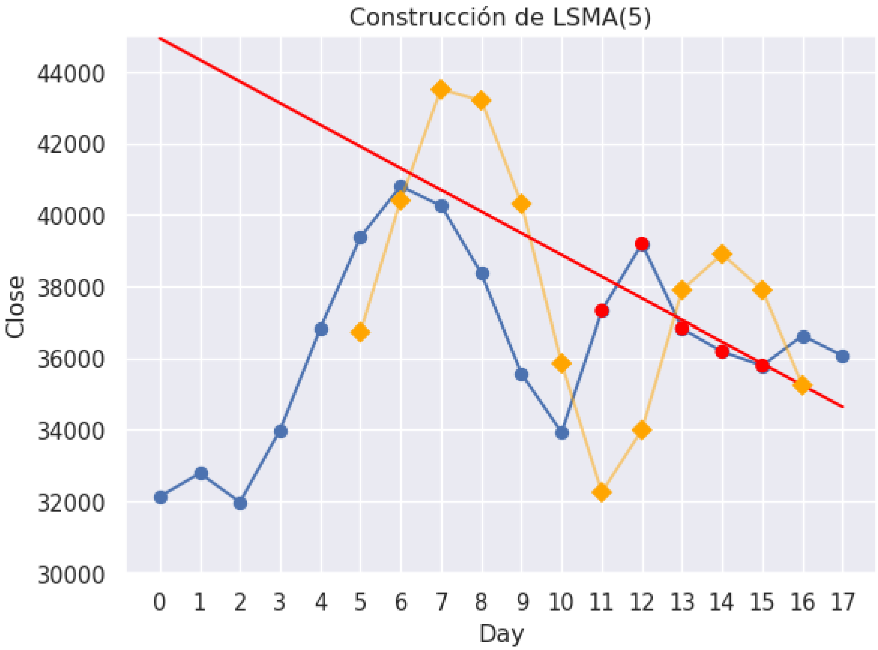

Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

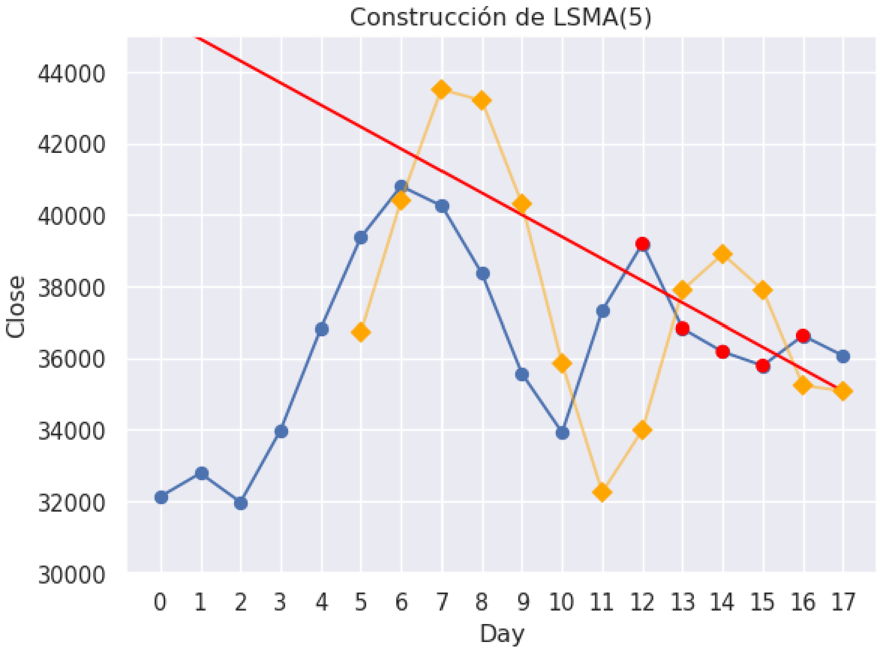

Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day

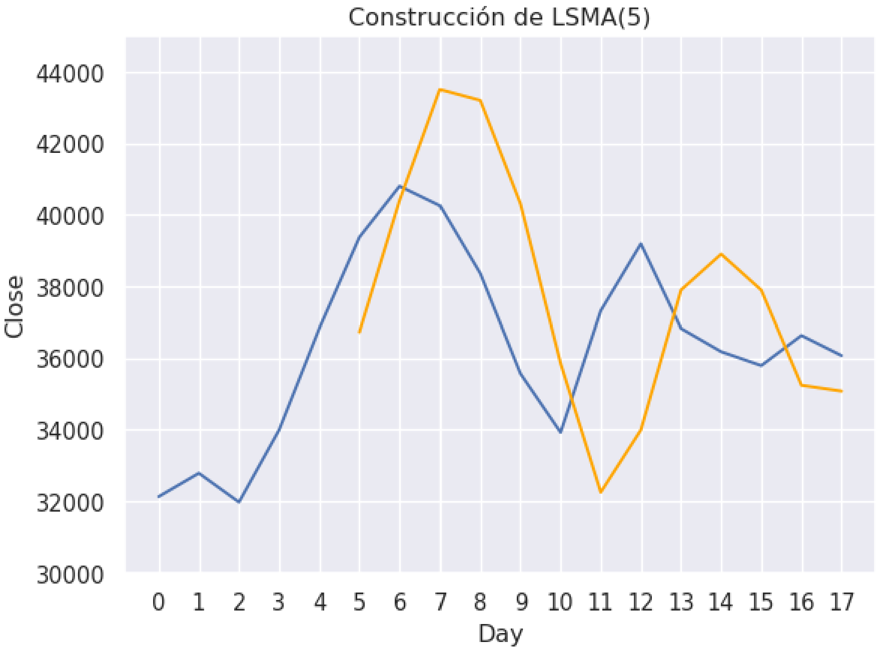

Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Day


Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17


Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17


Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
Day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17


Construcción de LSMA(5)
Close
Day
44000
42000
40000
38000
36000
34000
32000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17


Construcción de LSMA(5)
Close
44000
42000
40000
38000
36000
34000
32000
30000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Day