Содержание.

Чем хороша стационарность?

Классический пример - продолжите последовательность?

1, 2, 4, 8, 16, ?

Чем хороша стационарность?

Правильный ответ - 42!

Чем хороша стационарность?

Если ряд не постоянен во времени, мы не можем делать никаких осмысленных прогнозов.

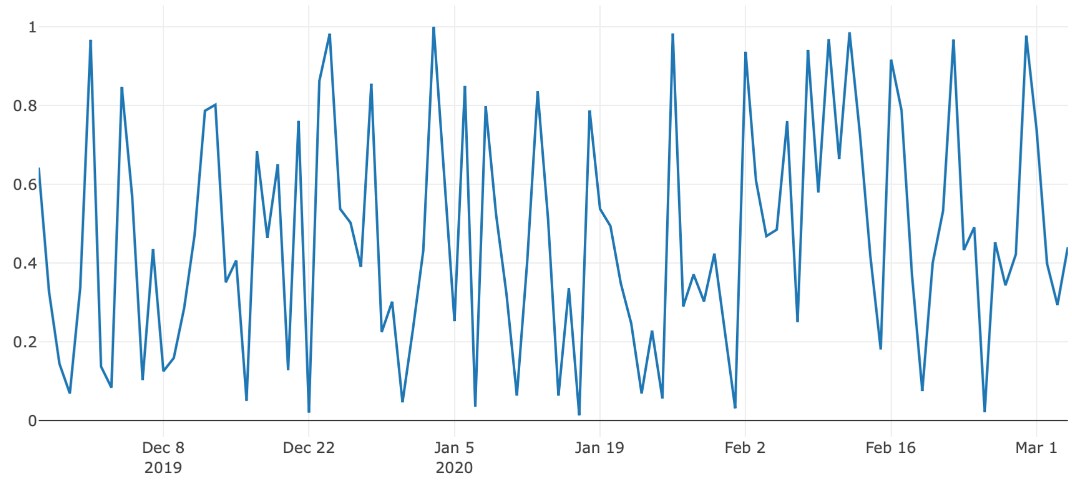
Стационарность дает нам гарантии того, что:

$$Var(Y_t) = const$$
$$E(Y_t) = const$$
$$Cov(Y_t, Y_{t-k}) = const$$

Какие бывают стационарные ряды

Самый простой тип - белый шум. $Y_t = \epsilon_t, \epsilon_t \sim N(0,\sigma^2)$

Всегда можем сказать, что этот ряд имеет матожидание 0 и дисперсию σ^2



MA(q) процесс

Следующий тип стационарный рядов - MA(q) процесс (Moving Average)

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

То есть каждая точка моделируется как линейная комбинация q предыдущих шумовых компонент

Большинство MA(q) процессов будут стационарными

*Примечание о шуме

Здесь и далее будет предполагаться, что шум это гауссов (нормально распределенный) шум с нулевым матожиданием и дисперсией σ^2

$$e_t \sim N(0, \sigma^2)$$

AR(p) - процесс

Данный процесс представляет собой зависимость каждый каждой точки ряда от P предыдущих точек

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Не любой AR процесс - стационарен

Для того, чтобы данный процесс был стационарен, необходимо выполнение следующего условия - необходимо, чтобы коэффициенты ϕ лежали на единичном круге, т.е., например

в AR(1) необходимо $-1 < \phi_1 < 1$;

в AR(2) необходимо $-1 < \phi_2 < 1,\ \phi_1 + \phi_2 < 1,\ \phi_2 - \phi_1 < 1$

ARMA процесс

Комбинация AR(p) и MA(q) процессов называется ARMA(p, q)

$$ARMA(p,q): \quad y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Истина - зачем нужна стационарность

По теореме Вольда - любой стационарный ряд с любой наперед заданной точностью может быть смоделирован моделью ARMA(p, q)!

Таким образом, сделав ряд стационарным, мы можем подобрать какой модели он соответствует.

Выбрать это модель для прогнозирования.

И восстановить исходный ряд обратными преобразованиями.

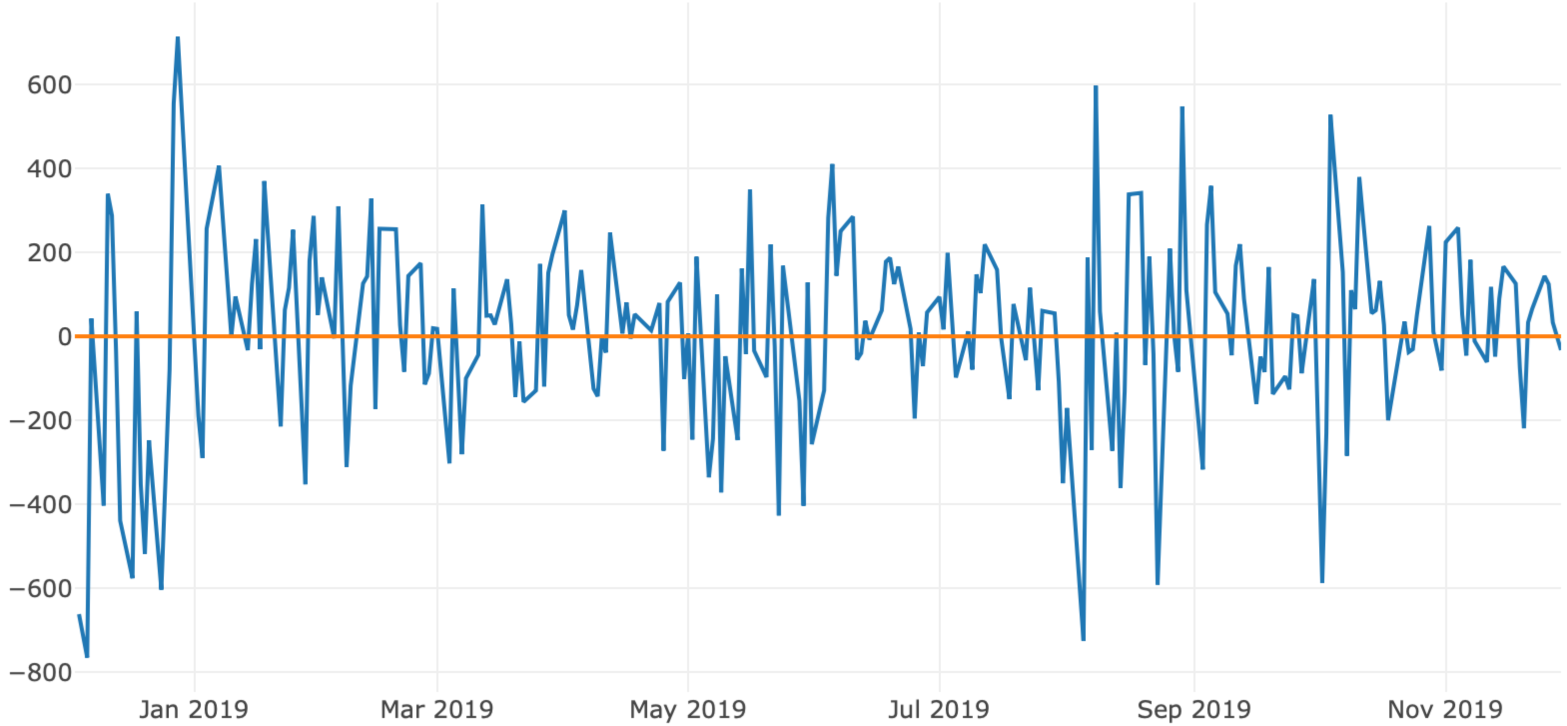
Как это правильно понимать?

Это что, мы теперь сможем предсказывать белый шум?

Нет, разумеется.

Теорема Вольда лишь говорит, что мы сможем понять какой модели ARMA(p, q) ряд соответствует. То есть, говоря простыми словами, для, например, белого шума, ответ будет таков: ваш ряд описывается моделью ARMA(0, 0).

Как это правильно понимать?



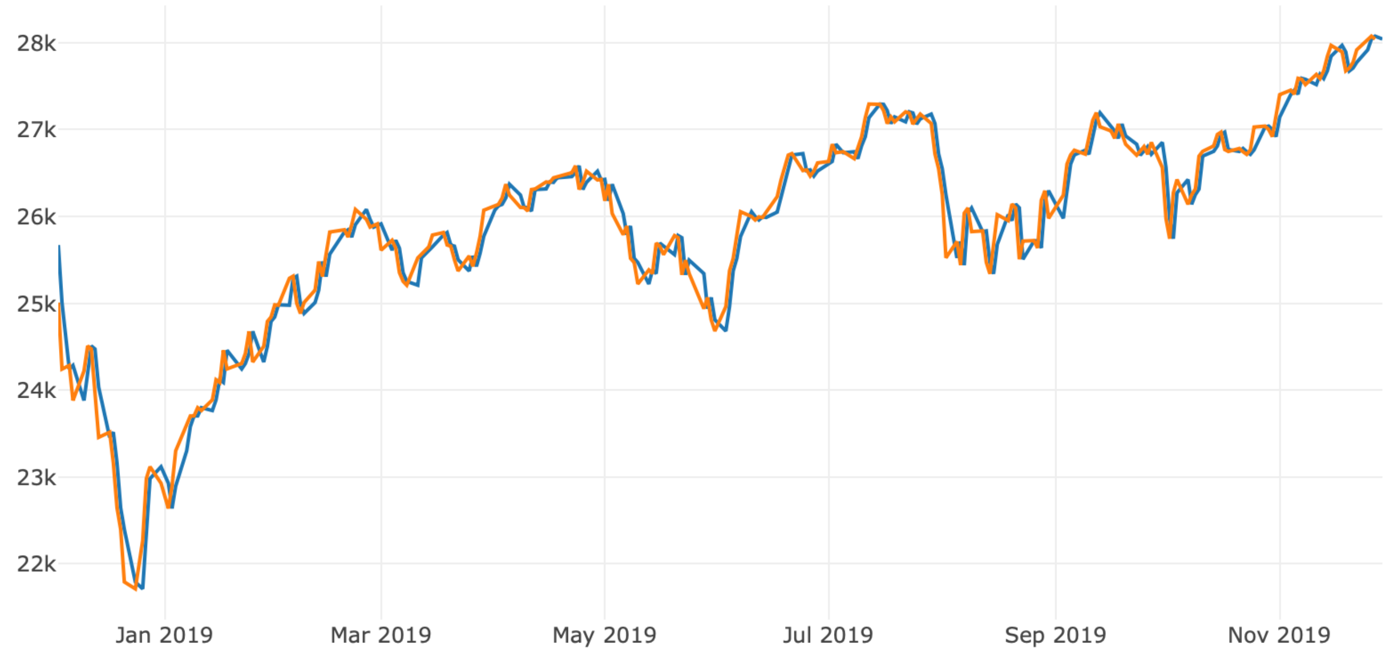
Модель ARIMA

Если ARMA работает для всех стационарных рядов, то почему не сделать модель, которая работает для всех рядов, что можно сделать стационарными, продифференцировав?

Ряд описывается моделью ARIMA(p, d, q), если d раз продифференцированный ряд описывается модель ARMA(p, q).

Как это правильно понимать?

Или индекс Доу-Джонса, допустим.



SARMA

Окей, моделью ARIMA мы можем смоделировать все стационарные ряды, а также все не стационарные которые можно сделать стационарными дифференцированием.

Проблема - сезонность не всегда можно убрать дифференцированием.

Решение - добавить в модель ARMA сезонные компоненты.

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

$$+ \phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \cdots + \phi_{PS} y_{t-PS}$$

$$+ \theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_{PS} \varepsilon_{t-QS}$$

Возьмем все вместе и получим SARIMA

Ряд описывается моделью SARIMA(p, d, q)(P, D, Q), если d раз обычно и D раз сезонно продифференцированный ряд описывается моделью SARMA(p, q)(P, Q)

$$y_t = \alpha + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

$$+ \phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \dots + \phi_{PS} y_{t-PS}$$

$$+ \theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \dots + \theta_{PS} \varepsilon_{t-QS}$$

Как подбирать параметры модели?

Как подобрать α, ϕ, θ (при условии зафиксированных гиперпараметров)?

По методу наименьших квадратов, так же как и в линейной регрессии.

Как подбирать параметры модели?

Порядки дифференцирования d, D?

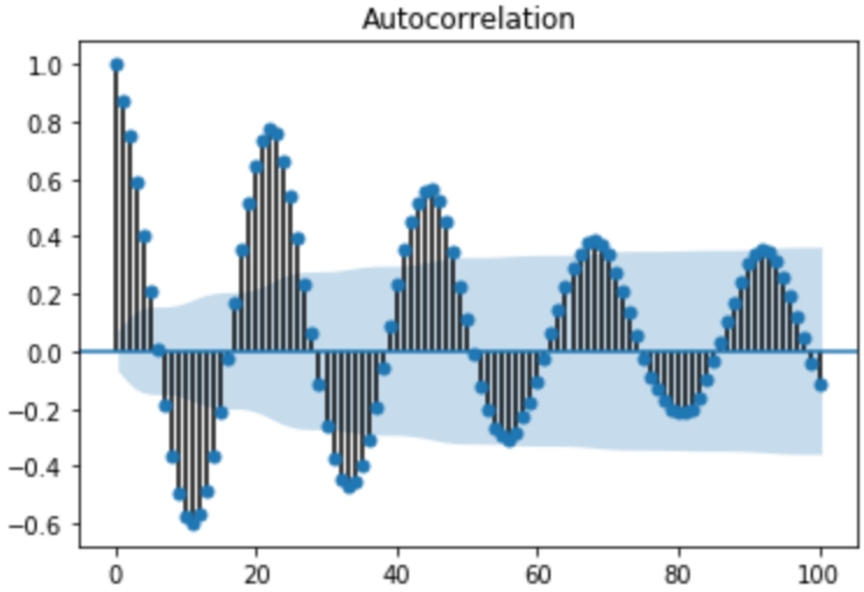
Пока ряд не станет стационарным (по критерию Дики-Фуллера, например)

Как подбирать параметры модели?

q, Q?

По графику автокорреляционной функции.

Q - последний значимый сезонный лаг
q - последний значимый не сезонный лаг

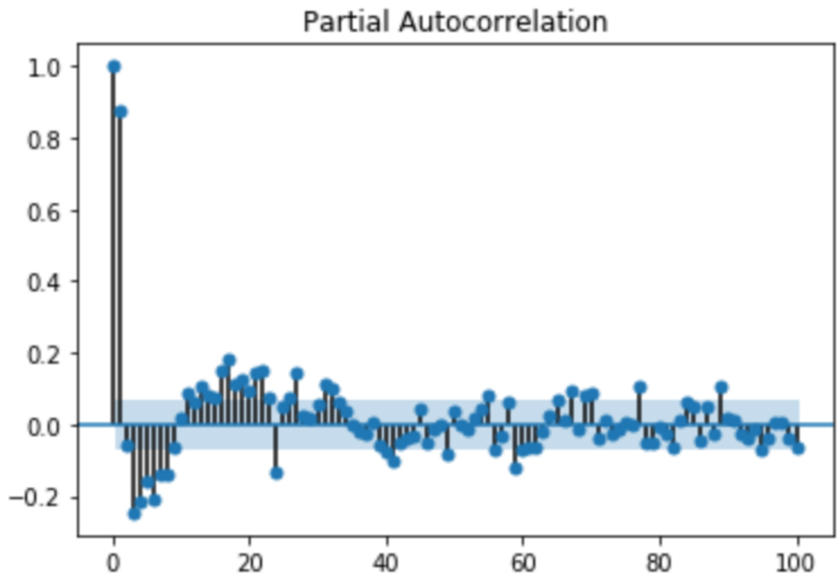


Как подбирать параметры модели?

p, P?

По графику частичной автокорреляционной функции.
*Частичная автокорреляция - автокорреляция после снятие регрессии на промежуточные значения.

p - последний значимый сезонный лаг
P - последний значимый не сезонный лаг



Критерий Акаике

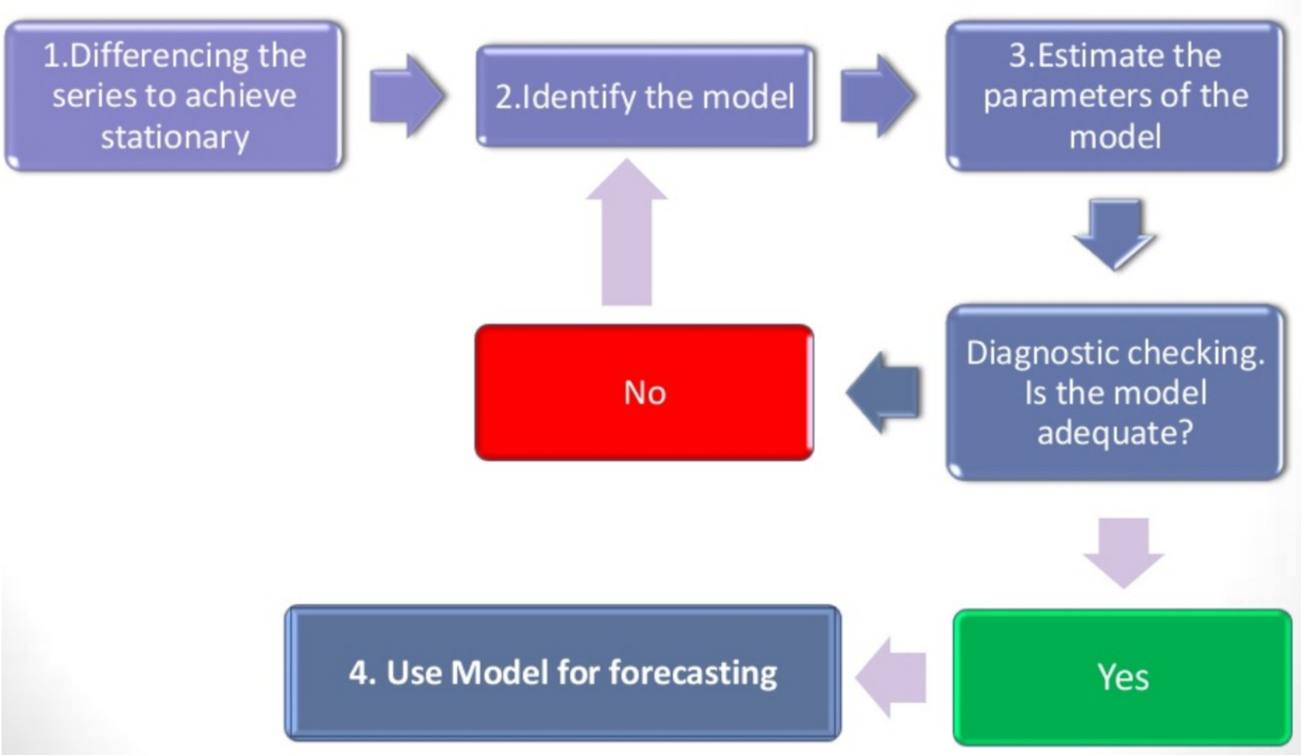
Выбор всех предыдущих параметров был приблизительным, что использовать для сравнения разных параметров?

Метод максимального правдоподобия не подходит, так как приводит к переобучению модели, просто выбирая максимальные значения параметров.

Поэтому используется информационный критерий Акаике.

$$AIC = 2k - 2\ln(L)$$

Алгоритм


1.Differencing the series to achieve stationary
2.Identify the model
3.Estimate the parameters of the model
Diagnostic checking. Is the model adequate?
No
Yes
4. Use Model for forecasting

Проблемы классического подхода.

Все вышеперечисленные подходы хороши, когда есть возможность “ручного управления”.

На практике в последнее время приходится иметь дело с огромным количеством рядов и настраивать параметры вручную - слишком затратно.

Можно использовать auto.arima, но ее возможности ограничены определенным типом модели.

Всегда только линейная регрессия.

Нет возможности добавить дополнительные признаки, вдруг я считаю, что популяция попугаев Какаду влияет на урожайность пшеницы в Ростовской области, как я могу включить это в модель?

Ответ - переход к задаче обучения с учителем.

1. Трансформировать временной ряд в матрицу обьекты-признаки.
2. Добавить дополнительные признаки по желанию.
3. Обучить произвольную модель или композицию моделей.
4. profit!

Матрица лагов.

2019-05-19 04:00:00	1.0
2019-05-19 05:00:00	6.0
2019-05-19 06:00:00	11.0
2019-05-19 07:00:00	16.0
2019-05-19 08:00:00	26.0
2019-05-19 09:00:00	38.0
2019-05-19 10:00:00	52.0
2019-05-19 11:00:00	81.0
2019-05-19 12:00:00	90.0
2019-05-19 13:00:00	25.0
2019-05-19 14:00:00	103.0
2019-05-19 15:00:00	87.0
2019-05-19 16:00:00	61.0
2019-05-19 17:00:00	48.0
2019-05-19 18:00:00	46.0
2019-05-19 19:00:00	33.0
2019-05-19 20:00:00	26.0
2019-05-19 21:00:00	8.0
2019-05-20 00:00:00	14.0
2019-05-20 01:00:00	6.0
2019-05-20 02:00:00	4.0
2019-05-20 03:00:00	14.0
2019-05-20 04:00:00	20.0
2019-05-20 05:00:00	30.0
2019-05-20 06:00:00	80.0

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0

Добавление произвольных признаков

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y	weekday	monthday	is_weekend	month	hour	mean	std
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0	6	19	0	5	11	21.428571	18.365340
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0	6	19	0	5	12	32.857143	26.585890
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0	6	19	0	5	13	44.857143	31.029172
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0	6	19	0	5	14	46.857143	28.858439
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0	6	19	0	5	15	59.285714	31.925509
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0	6	19	0	5	16	68.000000	29.563491
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0	6	19	0	5	17	71.285714	26.824829
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0	6	19	0	5	18	70.714286	27.341752
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0	6	19	0	5	19	65.714286	28.329692
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0	6	19	0	5	20	57.571429	28.377557
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0	6	19	0	5	21	57.714286	28.188143
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0	0	20	1	5	0	44.142857	25.491362
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0	0	20	1	5	1	33.714286	19.189531
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0	0	20	1	5	2	25.857143	17.324632
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0	0	20	1	5	3	19.571429	15.873008
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0	0	20	1	5	4	15.000000	10.785793
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0	0	20	1	5	5	13.142857	7.904188
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0	0	20	1	5	6	13.714286	9.050125

Обучение произвольной модели

- Linear regression
- Decision trees
- Gradient boosting decision trees
- Classical neural networks
- GRU, LSTM

Отличия от классической задачи регрессии

- Выбор параметров
- Метод кросс-валидации - никаких k-fold!