Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

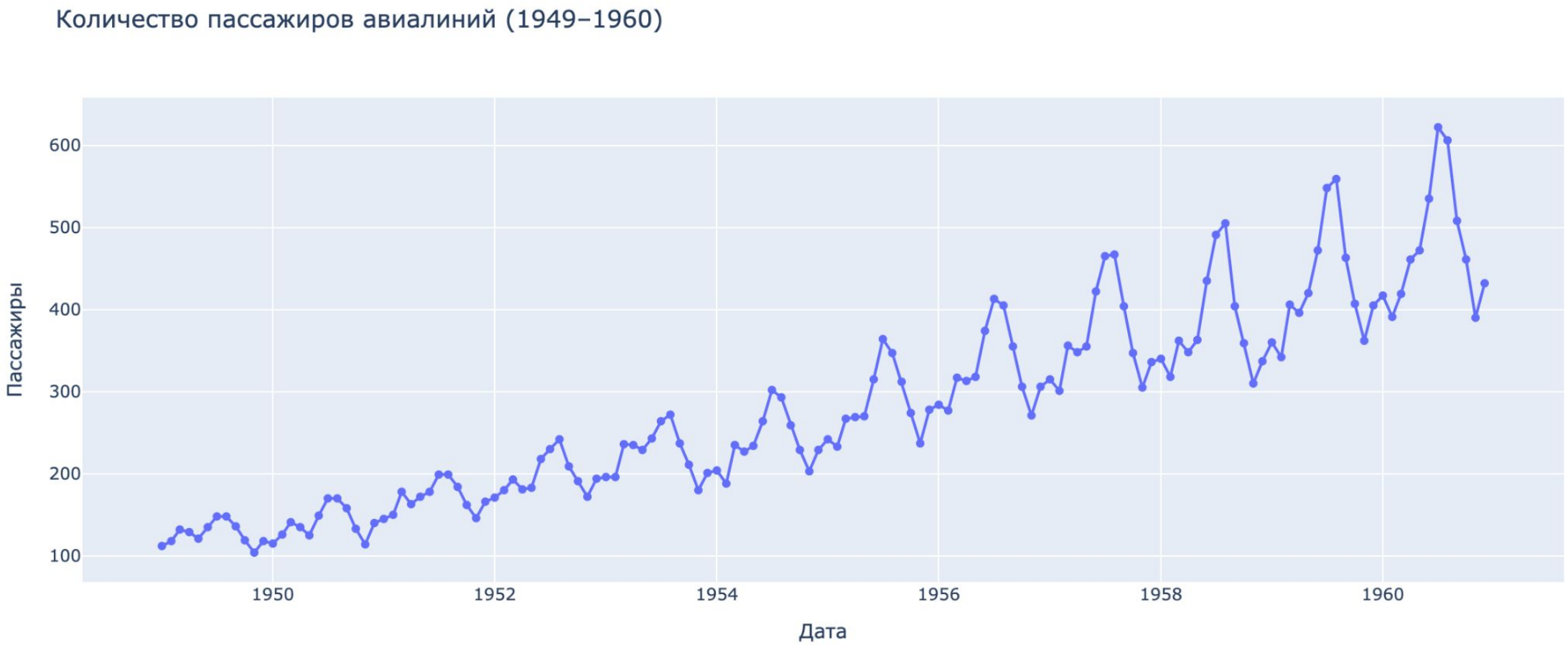
Лекция 12: Введение во временные ряды

Дата: 23.06.2025

Что такое временной ряд

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Временной ряд — это последовательность данных, измеренных в определенные моменты времени через равные промежутки (например, ежедневно, ежемесячно, ежегодно).



Что такое временной ряд.Формальное определение

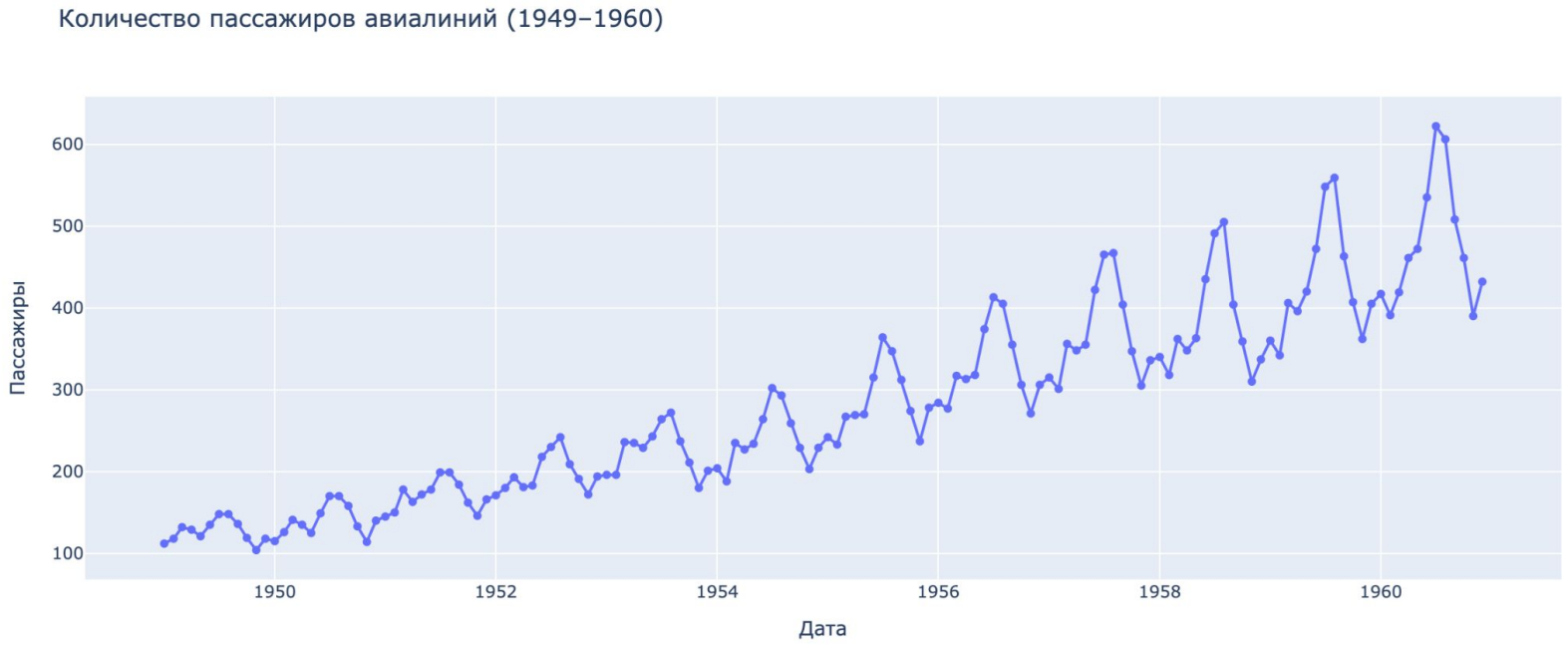
ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Временной ряд — это упорядоченная совокупность наблюдений $\{y_t\}$, где:

- t — индекс времени (например, $t = 1, 2, \ldots, T$),
- y_t — значение ряда в момент t.

Примеры временных рядов:

- Ежемесячные продажи магазина,
- Суточная температура воздуха,
- Биржевые котировки акций.



Работа с временными рядами

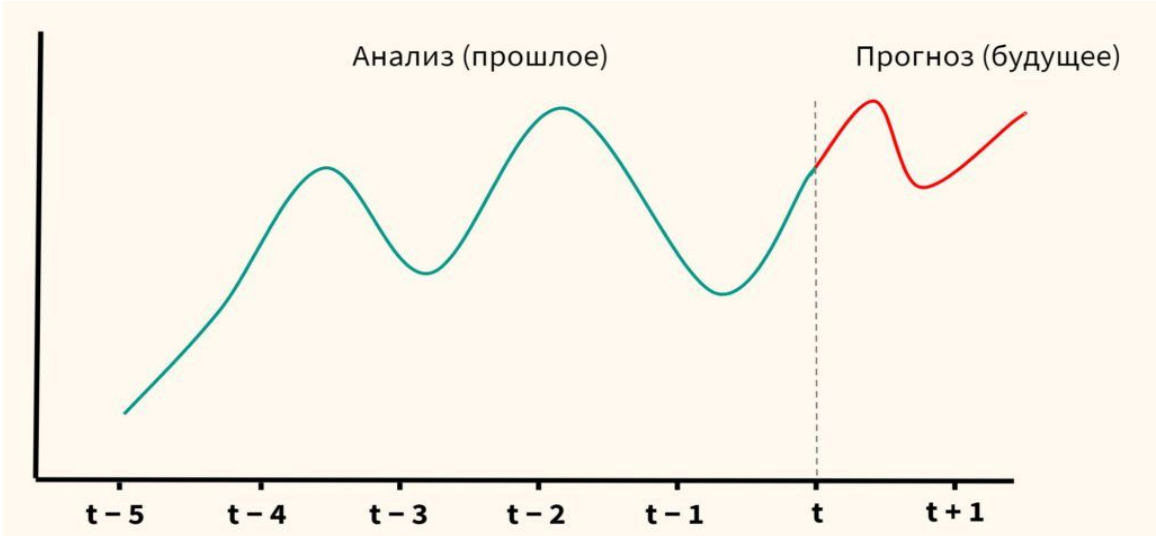
ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Анализ временного ряда (time series analysis)

Извлечение информации из данных
Заполнение пропусков
Обнаружение аномалий

Моделирование и прогноз (time series forecasting)

Прогнозирование будущих значений ряда по значениям в предыдущие моменты времени



Время t настоящее,
t−1, t−2,… прошлое, t+1, t+2,…

Шаг, сдвиг и скользящее среднее

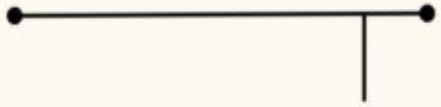
ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Изменить шаг (resample) - поменять временные периоды (группировка или разгруппировка данных)
Сдвиг (shift) на n периодов вперед или назад
Скользящее среднее (moving average, rolling average) за n предыдущих периодов (за окно, window). Скользящее среднее сглаживает временные показатели

Пример трехдневного скользящего среднего

Дни	1	2	3	4	5	6	7
Продажи, тыс. ед.	35	41	28	32	33	29	31

$(35 + 41 + 28) / 3 \approx 34{,}7$



$(41 + 28 + 32) / 3 \approx 33{,}7$

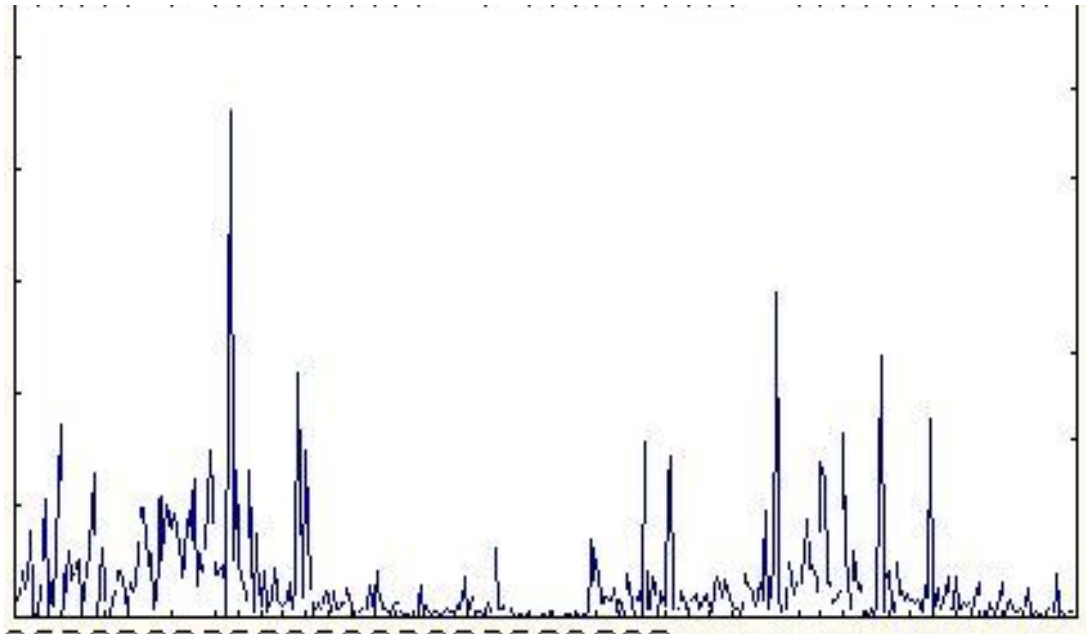
ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

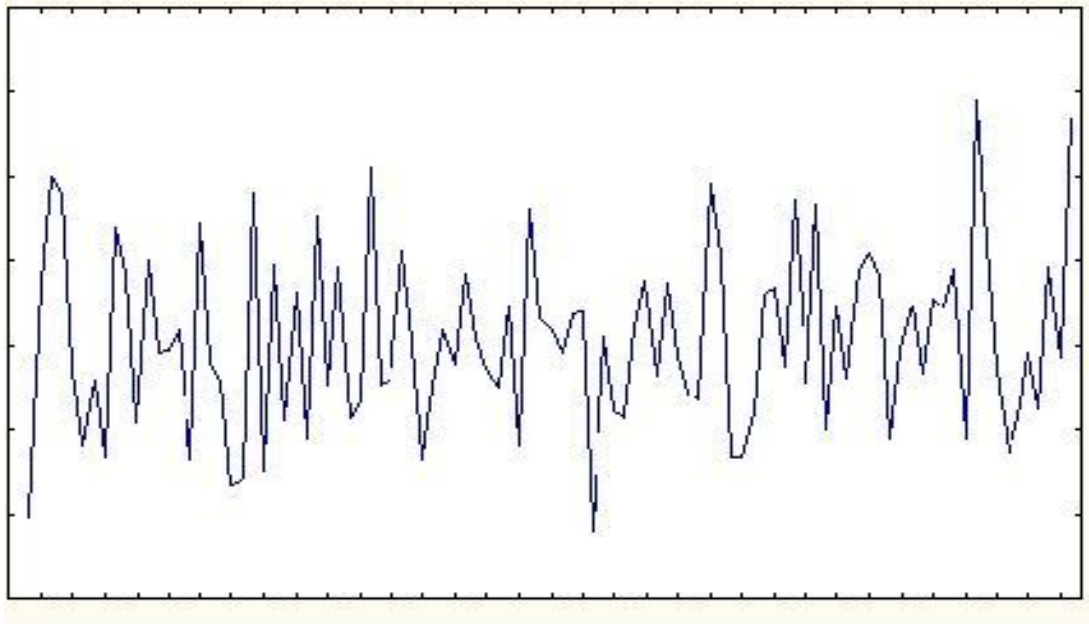
Демо как работать с данными, в которых есть дата

Стационарный ряд

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Стационарный ряд — это временной ряд, статистические свойства которого не зависят от времени. Это означает, что его основные характеристики (такие как среднее значение, дисперсия и автокорреляция) остаются постоянными на протяжении всего ряда.





Стационарный ряд. Тест Дики-Фуллера (ADF-тест)

Нулевая гипотеза H_0: Ряд нестационарен (есть единичный корень).

Альтернатива H_1: Ряд стационарен.

Как интерпретировать?

- Если **p-value < α (обычно 0.05)** → отвергаем H_0 → ряд стационарен.
- Если **p-value > α** → не отвергаем H_0 → ряд нестационарен.

Прогнозирование

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- AR (модель авторегрессии)
- MA (модель скользящего среднего)
- ARMA (Модель авторегрессии - скользящего среднего)
- ARIMA (Интегрированная Модель авторегрессии - скользящего среднего)
- SARIMA (Интегрированная Модель авторегрессии - скользящего среднего с учетом сезонности)
- ARIMAX, SARIMAX (X - eXtended) - возможность учета дополнительных внешних факторов

Фактор *— линейная суперпозиция переменных, которые сильно коррелируют между собой, при том что сами факторы не коррелируют*

AR (модель авторегрессии)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Метод авторегрессии моделирует следующий шаг в последовательности как линейную функцию наблюдений на предыдущих временных шагах

Регрессия ряда на собственные значения в прошлом

В модели **авторегрессии** мы прогнозируем интересующую переменную, используя линейную комбинацию прошлых значений переменной.

$$X_t = c + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \varepsilon_t$$

p - порядок модели

MA (модель скользящего среднего)

ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

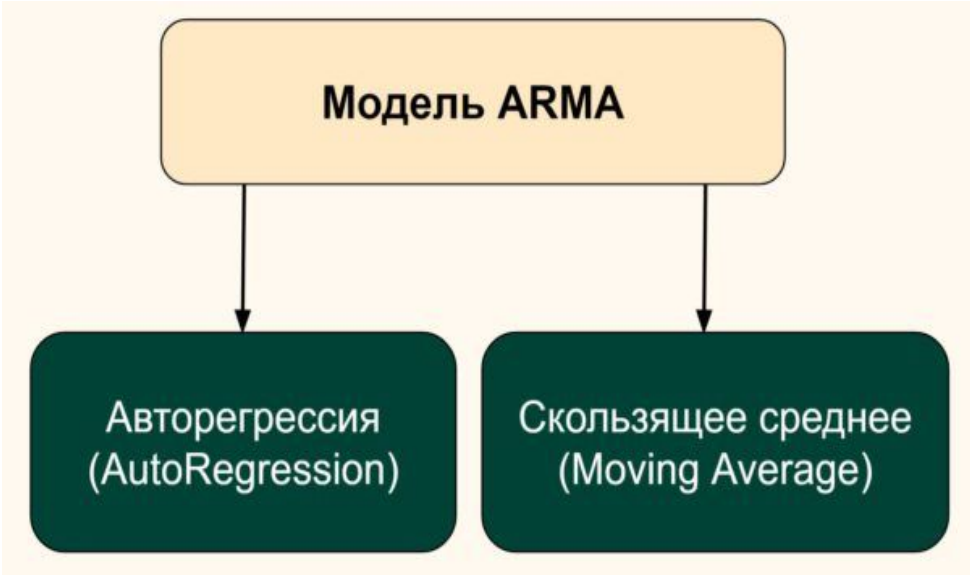
Метод скользящего среднего (MA) моделирует следующий шаг в последовательности как линейную функцию остаточных ошибок от процесса на предыдущих временных шагах

Модель скользящего среднего отличается от расчета скользящего среднего временного ряда.

$$X_t = \sum_{i=1}^{q} \theta_i \cdot \varepsilon_{t-i} + \varepsilon_t$$

ARMA (Модель авторегрессии - скользящего среднего)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС



Когда мы прогнозируем значение в период t с помощью данных за предыдущий период (AR(p), где p - сколько предыдущих периодов использовать)

$$y_t = c + \varphi \cdot y_{t-1}$$

где c — это константа, φ — вес модели, y_{t-1} — значение в период $t-1$

ARMA (Модель авторегрессии - скользящего среднего)

Метод авторегрессионного скользящего среднего моделирует следующий шаг в последовательности как линейную функцию наблюдений и ошибок на предыдущих временных шагах

Он сочетает в себе модели **авторегрессии (AR)** и **скользящего среднего (MA)**

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \cdot \varepsilon_{t-i}$$

ARIMA: AR(p)+I(d)+MA(q) = ARIMA(p,d,q)

ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Метод авторегрессионного интегрированного скользящего среднего (ARIMA) моделирует следующий шаг в последовательности как линейную функцию разностных наблюдений и остаточных ошибок на предыдущих временных шагах

Он сочетает в себе ***модели авторегрессии (AR)*** и ***скользящего среднего (MA)***, а также этап предварительной обработки разности, чтобы сделать последовательность стационарной

Метод подходит для временных рядов с трендом и без сезонных составляющих.

Добавляется компонент Integrated (I), который отвечает за удаление тренда (сам процесс называется дифференцированием)

ARIMA: AR(p)+I(d)+MA(q) = ARIMA(p,d,q)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

ARIMA представляет собой комбинацию трех моделей

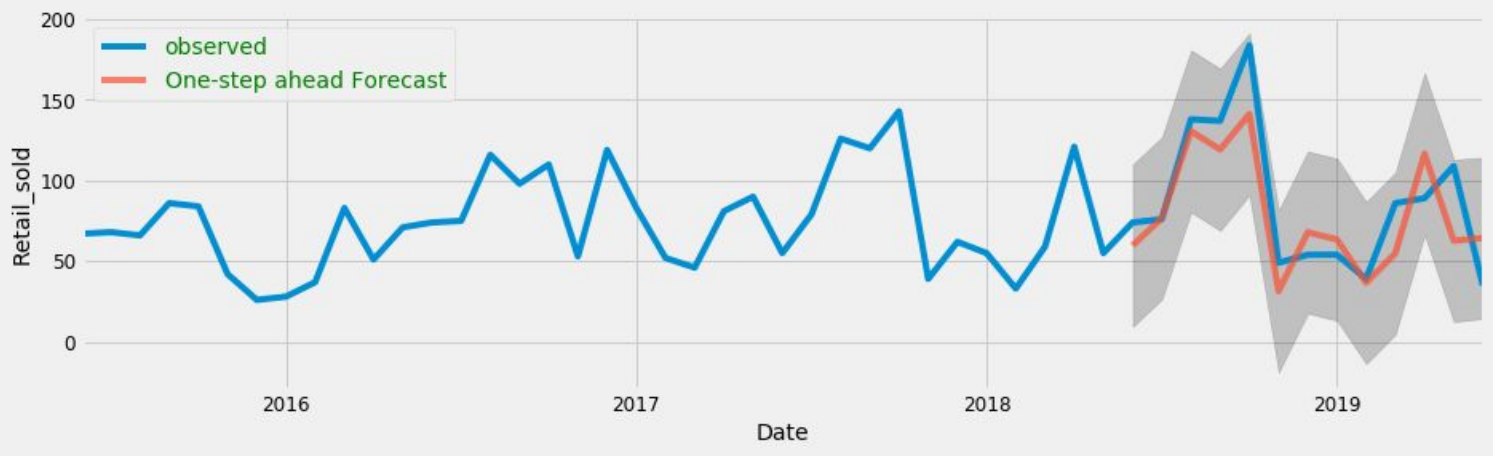
- AR (p) Авторегрессия - модель авторегрессии, которая использует зависимость между текущим наблюдением и наблюдениями за предыдущий период или периоды.
- I (d) Интеграция - использует разность наблюдений, чтобы сделать временной ряд стационарным.
- MA (q) Moving Average - модель, которая использует зависимость между наблюдением и остаточной ошибкой из модели скользящего среднего, применяемой к запаздывающим наблюдениям.

SARIMA (S - seasonal)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Модель имеет набор параметров:

- p, d, q - для модели ARIMA
- P, D, Q - для сезонности
- m - представляет количество точек данных (строк) в каждом сезонном цикле



Параметры модели можно подбирать с помощью auto_arima(). Выбор наиболее удачных параметров осуществляется на основе определенного критерия

Модель учитывает сезонность (Seasonality, S)

ARIMAX, SARIMAX (X — eXtended)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

SARIMAX включает еще и внешние или экзогенные факторы (eXogenous factors, отсюда и буква X в названии), которые напрямую не учитываются моделью, но влияют на нее.

Параметров у модели SARIMAX больше:

SARIMAX(p, d, q) x (P, D, Q, s)

- p и q, у нас появляется параметр d - отвечает за тренд
- набор параметров (P, D, Q, s) отвечает за сезонность.

SARIMAX: Introduction

$$y_t = \beta_t x_t + u_t$$

$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\epsilon_t$$

```
# Dataset
raw = pd.read_stata(StringIO(friedman2))
raw.index = raw.time
data = raw.ix['1981']

# Variables
endog = data.ix['1959':, 'consump']
exog = sm.add_constant(data.ix['1959':, 'm2'])
nobs = endog.shape[0]

# Fit the model
mod = sm.tsa.statespace.SARIMAX(
    endog.ix[:'1978-01-01'],
    exog=exog.ix[:'1978-01-01'],
    order=(1,0,1)
)
fit_res = mod.fit()
print fit_res.summary()
```

Statespace Model Results

Dep. Variable:	consump	No. Observations:	92
Model:	SARIMAX(1, 0, 1)	Log Likelihood	-340.508
Date:	Wed, 11 Feb 2015	AIC	691.015
Time:	10:29:10	BIC	703.624
Sample:	01-01-1959 - 10-01-1981	HQIC	696.105

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-36.0609	42.932	-0.840	0.401	-120.207	48.085
x1	1.1220	0.038	29.417	0.000	1.047	1.197
ar.L1	0.9348	0.054	17.385	0.000	0.829	1.040
ma.L1	0.3091	0.114	2.705	0.007	0.085	0.533
sigma2	93.2556	13.753	6.781	0.000	66.300	120.212

Метрики ARIMA, SARIMA

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- ***Средняя абсолютная ошибка (Mean Absolute Error, MAE)*** - это степень несоответствия между фактическими и прогнозируемыми значениями.

- ***Среднеквадратическая ошибка (Mean Squared Error, MSE)*** измеряет среднюю квадратическую разницу между оценочными значениями и фактическим значением. Чем ниже значение MSE, тем лучше модель способна точно предсказывать значения.

- ***Квадратный корень из MSE (Root Mean Squared Error, RMSE)*** для того, чтобы показатель эффективности MSE имел размерность исходных данных, из него извлекают квадратный корень и получают показатель эффективности RMSE

Метрики ARIMA, SARIMA

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- ***Информационный критерий Акаике (AIC)*** полезен при выборе предикторов для регрессии, также полезен для определения порядка построения модели ARIMA. Критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных. Существует также AICC

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

где k — число параметров модели, L — максимизированное значение функции правдоподобия модели. Лучшей признается та модель, для которой значение AIC минимально.

- ***Байесовский информационный критерий (Bayesian information criterion - BIC).*** Критерий основан на использовании функции правдоподобия и тесно связан с информационным критерием Акаике

$$BIC = k \cdot \ln(n) - 2\ln(\widehat{L}),$$

где $\widehat{L}$ — максимальное значение функции правдоподобия наблюдаемой выборки с известным числом параметров, k — число параметров модели, n — объем обучающей выборки.

Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Спасибо за внимание