Модуль 2 Методы Предсказания ВР.

**Введение:**

Модуль дает представление об основные понятия о ВР и методах его осмысления. В том числе:

* базовые методы предварительного анализа ВР,
* типы решаемых задач анализа ВР
* модели ВР
* Основные статистические свойства ВР
* Понятие о особых свойствах ВР, например:
  + АКФ,
  + Гетеросекдостичность
  + Белый Гауссов Шум,
  + особенности стационарного вида ВР.
* Виды простых предсказаний ВР, в т.ч.:
  + Наивные методы предсказаний,
  + Методы на основе скользящего среднего и экспоненциального сглаживания,
  + Особенности использования методов линейной регрессии в анализе ВР
  + Особенности использования методов нелинейной регрессии в анализе ВР
  + Особенности базовых методов разложения ВР

Также Вы познакомитесь с методами работы с ВР в рамках языка программирования Python, включая особенности:

* визуализации ВР
* определения базовых свойств ВР
* моделирования ВР
* особенности работы с фреймворком SKTime
* особенности выбора методов предсказания ВР

В результате прохождения модуля вы будете:

* Знать Основные понятия анализ временных рядов, основные свойства ВР
* Уметь предложить подход для решения той или иной задачи анализа временных рядов.
* Владеть методами предварительного анализа временных рядов

Модуль состоит из 2 лекций и 3 практик.

Лекции

1. Введение во ВР
2. Лекция 2 - простые методы предсказания ВР

Практики

1. Предварительный анализ ВР
2. Моделирование ВР
3. Знакомство с библиотекой SKTime и простые методы предсказания ВР

**В ходе изучения модуля основными понятиями явлются:**

**Модель авторегрессии - скользящего среднего (АРСС, ARMA) –** это модель которую можно проинтерпретировать как: текущее значение ВР зависит от прошлых значений до лага p (**AR часть**) и от текущего и прошлых внешних «возмущений» (флуктуаций) до лага q (**MA часть**). Это параметрическая модель c двумя параметрами (p,q).

**Задача ARMA-аппроксимации** (построения АРСС модели) - найти весовые коэффициенты и и оценить их порядки (p,q), приближающие лучше всего к исходному процессу

**Лаг модели** – это значение задержки рассматриваемых отсчетов относительно заданного 0.

**Начальные значения порядков ARMA моделей** могут быть выбраны из анализа графиков автокорреляционной функции (АКФ, ACF) – MA часть и частичной АКФ (ЧАКФ, PACF) – AR часть как значения лагов перед резким спадом к нулю в обоих случаях. Однако это справедливо только для стационарной в слабом смысле модели. Тестирование Лагов АКФ на значимость выполняется при помощи статистических тестов.

**Дообучение** **ARMA моделей** заключается в настройке их порядков и соответствующих им коэффициентов. Оно осуществляется при помощи информационных критериев и последующего тестирования модели. При этом чем ниже общий порядок модели – тем ниже вероятность переобучения.

**Информационный критерий** – критерий, учитывающий, как дисперсию остаточной части модели, так и порядки модели. Критерий имеет исключительно смысл в ранжировании (используется только для сравнения моделей, чем меньше значение, тем лучше). Критерий позволяет задать мини-максуную задачу поиска параметров моделей: минимум ошибки при максимально-допустимом числе параметров модели.

**Модель называется интегрированной авторегрессионной скользящей средней (ARIMA) –** модель нативно учитывающая численное дифференцирование для достижения стационарности результат моделирования.

Модель **ARIMA** имеет три порядка (p,d,q), где **d – порядок численной производной** (разности). Модель удобно записать в лаговой форме. Важно понимать, что дифференцированию подлежит только AR часть. В первую очередь такая модель может компенсировать нестационарность тренда или другие медленные изменения. При выборе порядков модели сначала должен быть найден порядок дифференцирования.

Для проверки модели на стационарность используются ряд визуальных и статистических тестов. Среди них:

* Статистика скользящего окна (Rolling Statistics)
* ACF анализ
* Расширенный тест Дики-Фуллера (Augmented Dickey-Fuller Test, ADF)
* Другие тесты, например Тест Квятковского – Филлипса – Шмидта – Шина

**Модель сезонной интегрированной авторегресси - скользящего среднего (SARIMA) -** это модель нативноучитывающая даже сравнительно быстрые и интенсивные, но регулярные нестационарности типа сезонность. Для этого испльзуется прием сезонного дифференцирования.

Модель имеет задание порядков SARIMA(p,d,q)×(P,D,Q)s, включающая как обычные p,d,q, так и сезонные составляющие P,D,Q для сезонности с периодом s.

При выборе начальных значений порядков модели следует сначала определить порядки обычной и сезонной производной, приводящие ВР к стационарному виду.

**Экзогенные факторы (exdogs, экзогенные ковариаты)**  или набор факторов () являются дополнительными факторам к целевой переменной , которые статистически не зависят от , но влияют на нее. Учет таких факторов может быть произведен в моделях SARIMA, такая форма называется SARIMAX.

**Комбинированные методы предсказания –** представляют собой попытку описания моделей несколькими последовательными составляющим таким образом чтобы остаток описания (необъясненная часть) была как можно меньше. Примерами подхода являются использование регрессии для моделированя тренда и моделей ARIMA для остатка такого подхода. Другими примерами такого подхода являются такие модели, как:

* **Модели BATS -** Box-Cox transformation, ARMA errors, Trend, and Seasonal components
* **Модели TBATS –** учитывающие нестационарность сезонности путем ее описания рядами Фурье
* Обобщенная Авторегрессионная условная гетероскедастичность, или **GARCH** – модель учитвающая модель ARCH для моделирования валатильности (явно моделирует изменение дисперсии) ВР и модель ARMA для оставшейся части
* **AR-NET** - модель представления AR моделей в виде аналога нейронной сети и использование продвинутых методов оптимизации в ней. А также модель **Neural Prophet** учитывающая AR-NET как одну из составляющих.

**Разведывательный анализ данных (EDA) -** относится к процессу выполнения начальных исследований над данными с целью очистки ВР, выдвижения гипотез о поведении ВР и его необходимых преобразованиях. К примерам EDA относятся:

* Визуализация ВР
* Анализ статистических свойств ВР
* Поиск и выявления неруглярностей или других подозрительных составляющих ВР
* Простые преобразования ВР
* Поредение основных свойств ВР: сезонность, тренд и тд
* И другие приемы призванные попытаться выявить основные сущности рассматриваемых данных

**Представление временного ряда** включаетрядприемовпо приведению ВР к виду, оптимальному для использования тех или иных методов решению задачи для него. К примерам такого представления относятся:

* аггрегация (укрупнение шагов, преобразование в несколько рядов),
* тренд-сезонная декомпозиция,
* Спектральный анализ и фильтрация (денойз например),
* декомпозиция по внутренним модам,
* Приближение/фильтрация через модели ARIMA/PCA и т.д.
* Представленя через лаги (АКФ), траекторные матрицы, другие табличные формы
* Преобразования (напр. на основе автокодирующей сети, вейвлет-представление),

**Заключение:**

Модуль «Основные понятие анализа Временных рядов» является важной частью курса, однако, показывающей лишь только основные – наиболее простые подходы к анализу этой модальности данных. Изученные материалы будут полезны при постановке и проверке базовых гипотез об анализируемом ВР и в рамках поставленных задач.

Изученный программные модули SKTime, а также особые модули Pandas, пожалуй, являются на сегодня практически базовыми фреймворками для анализа ВР различного уровня сложности.

Теоретические материалы, изученные в модуле, носят характер дискурса и раскрывают основные нарративы изучаемой дисциплины. В т.ч. модуль дает понимание основных концепций, такие как:

* Задачи анализа ВР,
* Тренд ВР,
* сезонность,
* представление о статистическом анализе ВР,
* анализ остаточных частей ВР,
* методы предсказания ВР.

Все эти концепции, хотя и могут быть переосмыслены в рамках иных подходов, объединившись создают область знаний, дающую передовые на сегодня методы решения релевантных задач.

В следующих модулях вы узнаете о более сложных подходах к анализу ВР, однако изученные в данном модуле технологии останутся базовыми (т.н. baseline), не позволяющим у ходить слишком далеко в априорных-синтетических суждениях более высокого порядка.