Лабораторная 2 - timeseries

[github classroom task](https://classroom.github.com/a/cTqy61bn)

Порядок сдачи и защиты лабы описан [тут](https://docs.google.com/document/d/1pj0LIO8kuOzqxgxW-dfZkQFhwZ9dkQ1hIxg5Va39x7U/edit?usp=drive_link)

Дедлайн указан в таблице с баллами

# Задание

Обучите две модели предсказания временного ряда и сравните их.

Нельзя использовать модели, **код которых** был разобран на лекции. Например:

* экспоненциальное сглаживание   
  простое, двойное, тройное (Хольта-Уинтерса)
* скользящее среднее  
  MA, ARMA

Принимать 350 лабораторных по SARIMA неинтересно. Поэтому за выбор другой модели засчитывается плюс в карму

Первую модель возьмите из библиотеки (например из рекомендованных ниже библиотек). В качестве второй модели реализуйте первую самостоятельно. Нужно реализовать и саму модель, и процесс её обучения. Метрики качества моделей должны быть одинаковыми или иметь пренебрежимо малую разницу, допустим 5% MAPE

Можно не реализовывать оптимизацию самостоятельно. Например, взять готовую линейную регрессию в качестве метода оптимизации параметров модели

## Данные

Если собранный вами датасет содержит предсказуемый временной ряд, то используйте его. Будем считать предсказуемым такой ряд, в котором больше 100 значений и значение ряда в 90+% случаев превышает 5. Чем больше значения и длительности, тем лучше

Например, если у вас есть датасет объявлений на авито, то можно создать временной ряд из количества объявлений, созданных в эту дату. Тогда если у вас есть данные за последовательные 100 дней, и в 90+% дней объявлений больше 5, то датасет отлично подходит для предсказания. Если это не выполняется, то вряд ли получится предсказать ряд с удовлетворительным качеством

В ином случае, можете предсказать цену акции интересной вам компании или актива. Или любой другой интересный вам ряд. Оригинальность засчитывается в вашу карму

## Предобработка

Временной ряд сложно предсказывать по дням. Агрегируйте ряд по неделям/месяцам/другим большим промежуткам. Длина итогового ряда должна быть больше 15 элементов

Удалите из ряда тренд, если он есть. Используйте другие рассмотренные на лекции приёмы

## Обучение

Обучите модели на 80% временного ряда. Предскажите оставшиеся 20% (как минимум 3 точки) ряда с помощью моделей. Отобразите на одном графике тренировочную, тестовую и предсказанную моделью части ряда

Посчитайте метрики качества модели. Метрика должна отражать разницу между предсказанными и истинными данными. Используйте рассмотренные на лекции метрики

# Рекомендации

* длина ряда от 500 значений
* количество 90+% значений больше 10
* **не стесняйтесь задавать вопросы в чат курса**
* не выбирайте arima и более простые модели - это скучно
* не выбирайте нейросетевые модели, если не понимаете на 101% как они работают и обучаются

Для реализации используйте python библиотеки

* statsmodels
* Prophet
* sklearn
* etna

# Популярные замечания

* используйте конфиг файл для констант
* не коммитить .idea
* функции длиннее 20 строк - плохо. Их можно декомпозировать
* некоторые модели работают только со стационарными рядами
* нужно нарисовать предсказание обеих моделей и истинные данные на одном графике
* нужно посчитать метрику качества предсказаний обеих моделей. Желательно легко интерпретируемую, например MAPE, SMAPE, MAE и тд
* если значения ряда большие (>30), то сильно поможет логарифмирование. Но не забудьте сделать обратное преобразование предсказаний

# Теормин

В вопросах под термином “ряд” подразумевается временной ряд

1. Что такое стационарность ряда, как проверить
2. Какие есть компоненты ряда, что из них можно узнать о ряде и как их рассчитывать
3. Как избежать переобучения при обучении ML моделей на рядах. Назовите хотя бы 3 способа
4. Что такое автокорреляция, как считать
5. Стационарные, интегрированные, сезонные, циклические и детерминированные ряды, примеры
6. Из чего состоит коррелограмма и что из неё можно узнать о ряде, как считается граница, в пределах которой корреляция статистически значима
7. Зачем нужен тест Дики-Фуллера, как он рассчитывается и что такое p-value. Какой статистический тест в нем используется и на каких значениях проходит тест (на значениях ряда или на средних значениях окон ряда или на чем-то другом?)
8. Зачем нужно дифференцирование ряда и как его проводить, почему называется дифференцированием
9. Зачем нужно логарифмирование ряда и как его проводить, чем отличается от преобразования Бокса-Кокса
10. Зачем нужно преобразование Бокса-Кокса и как оно рассчитывается, как выбрать оптимальный параметр для него
11. Метод экспоненциального сглаживания. Как считать, плюсы и минусы
12. Модель Хольта-Уинтерса. Как считать, плюсы и минусы
13. Модели AR и ARMA. Как считать, плюсы и минусы
14. Модель ARIMA. Как считать, плюсы и минусы, ограничения
15. Модель SARIMA. Как считать, плюсы и минусы, ограничения
16. Какие метрики используются для оценки качества моделей рядов, хотя бы 3
17. Как выбрать оптимальное количество lag-ов для модели ARIMA
18. Как прогнозировать многомерные ряды, модель VARIMА, как она считается
19. Вопрос-приз: Можно ли использовать нейронные сети для анализа и предсказания рядов
20. Как обрабатывать пропущенные значения в рядах. Назовите хотя бы 3 способа

# FAQ

1. можно ли использовать np.linalg и другие математические операции?  
   Да, можно. Вам нужно реализовать именно модель - ее параметры и их настройку
2. Eсли в датасете, который парсил ранее, нет временного ряда, можно ли взять датасет из открытых источников?  
   Да, можно. В разделе Данные так и написано
3. Можно ли использовать для первой модели другие библиотеки, например pmdarima?   
   Да, список библиотек - это рекомендация. Используйте что хотите
4. Длина вычисленного seasonal decompose тренда, который является компонентой ряда, меньше, чем сам ряд. Как вычесть тренд из ряда?  
   Хороший путь: посчитать тренд самостоятельно, используя паддинг  
   Плохой путь: убрать начало ряда, на которое не хватает тренда
5. На реализацию в лабе остается только SARIMA. Что еще можно реализовать? Деревья решений?  
   Нужно реализовать алгоритм предсказания рядов. Деревья - алгоритм классификации. В рекомендованных выше библиотеках вы найдете алгоритмы предсказания рядов. Некоторые проще, чем SARIMA, некоторые интереснее