Задание для обсуждения на следующее занятие

Задача: Прогнозирование уровня средней заработной платы в России

Для выполнения этого задания вам понадобятся данные о среднемесячных уровнях заработной платы в России с января 1993 по 2020. Дописать в конец ряда данные за следующие месяцы, если они уже опубликованы; найти эти данные можно, например, здесь.

Необходимо проанализировать данные, подобрать для них оптимальную прогнозирующую модель

1) в классе ARIMA и построить прогноз на каждый месяц на два года вперёд от конца данных.

Придерживайтесь стандартного алгоритма построения прогноза:

a) Визуальный анализ ряда Дайте ответ на вопросы.

   Используя метод скользящего среднего, сгладьте временной ряд. Ширину окна выбирайте равной одной недели, месяцу, кварталу. На одной графике изобразите исходный временной ряд, а также его сглаженные версии. Есть ли у ряда тренд, сезонность? Является ли заданный временной ряд стационарным? Почему? Проверьте своё предположение с помощью теста Дики-Фуллера.

b) Стабилизация дисперсии (при необходимости)

c) Выбор порядка дифференцирования

d) Выбор начальных приближений для p, q, P, $\mathrm{Q}p,q,P,Q$

e) Обучение и сравнение моделей-кандидатов, выбор победителя

f) Анализ остатков построенной модели, при необходимости — её модификация: если остатки получаются смещёнными, прогноз нужно скорректировать на константу; если остатки нестационарны или автокоррелированы, можно попробовать расширить область, в которой подбираются значения p, q, P, $\mathrm{Q}p,q,P,Q$.

g) Прогнозирование с помощью финальной модели.

2) Используя методы машинного обучения

   a) Создайте матрицу признаков для заданного временного ряда.

   b) Сравните производительность следующих алгоритмов, используя кросс-валидацию:

   c) `SGDRegressor`

d) `LinearRegression`
e) `RandomForestRegressor`
f) `XGBRegressor/LGBMRegressor`
g) Обучите бустинг-модель и постройте график важности признаков (`feature_importances`).
h) Разбейте временной ряд на тренировочный и тестовый так, чтобы в тестовой части были данные за 2 последних месяца. Обучите по одной линейной и древесной модели, которые показали себя лучше всего в 3-м пункте, и спрогнозируйте значения для тестовых данных и тренировочных данных.
i) На одном графике изобразите исходный временной ряд и предсказания всех моделей, построенных в предыдущем пункте. Используйте разные цвета для истинных и предсказанных значений. Прогнозы разных моделей также изображайте разными цветами. Добавьте на график легенду. Выделите участок, на котором изображаются предсказания для тестовых данных (например, сделайте фон темнее).
j) Попробуйте улучшить работу алгоритмов, подбирая параметры моделей или изменяя исходный ряд, например, при помощи преобразования Бокса-Кокса.

3) Используя методы сглаживания.

Подготовить ipython-ноутбук с проведённым анализом; пожалуйста, комментируйте в ноутбуке каждый важный шаг построения модели. Сделайте вывод по всем моделям.