Направление Data Scientist, компания McKinsey

Добро пожаловать на виртуальную стажировку компании McKinsey!

Предлагаем тебе примерить роль специалиста по данным (Data Scientist) в международной консалтинговой компании, где решаются сложные задачи для крупнейших предприятий по всему миру.

Направления Digital и Analytics активно развиваются в McKinsey. При этом отличительной особенностью Data Science в консалтинге является возможность попробовать себя в различных индустриях: банкинге, транспортном и энергетическом секторах, промышленности, FMCG[1]. Тебе придется сопровождать проекты от идеи до реализации и работать в кросс-функциональных командах, что потребует сильных soft skills и глубокого понимания бизнес-процессов.

Мы рекомендуем выполнять задания в указанном порядке, так как они взаимосвязаны между собой и объединены общей темой анализа ежедневного потребления и производства «зеленой» энергии в Австрии за период с 2015 по 2020 год.

Выполнение всего блока заданий займет у тебя не более 60-80 минут.

По результатам выполнения заданий ты научишься:

1. Анализировать сезонность данных, работая с временными рядами и используя Python, а также визуализировать данные с помощью построения различных графиков.
2. Использовать Ensemble Machine Learning (ML) алгоритмы, в частности Decision Tree + AdaBoost, для прогнозирования временных рядов.
3. Выдвигать бизнес-гипотезы на основе анализа данных, а также рассчитывать их потенциальный эффект.

Рекомендуемый тайминг:

1. 15-20 минут на первое задание.
2. 30 минут на второе задание.
3. 30 минут на третье задание.

Информация о загрузке решения:

Данный проект содержит несколько подзадач. Можно загрузить файл, содержащий решение только части заданий, но по возможности старайся сделать их все.

Желаем удачи!

[1] FMCG (fast –moving consumer goods) — товары повседневного спроса, включающие продукты легкой и пищевой промышленности, а также косметику, предметы личной гигиены, моющие средства и пр.

Дата-сет

Для выполнения заданий виртуальной стажировки предлагаем воспользоваться открытым дата-сетом, сгенерированным на основе информации Open Power System Data (OPSD) для Австрии: **opsd_austria_daily.csv**. Основными источниками данных являются различные европейские операторы систем передачи (TSO).
Этот набор данных содержит значения ежедневного потребления энергии и ее производства с помощью ветровых установок и солнечных панелей в течение 2015–2020 годов и даты:

- **Date**: Дата в формате year-month-day.
- **Electricity_consumption**: Потребление электричества в ГВт·ч.
- **Wind_production**: Производство ветровой энергии в ГВт·ч.
- **Solar_production**: Производство солнечной энергии в ГВт·ч.
- **Price**: Спотовая цена[2] на электроэнергию для Австрии в евро за 1 кВт·ч.
- **Wind+Solar**: Сумма ветровой и солнечной энергии в ГВт·ч.

[2] Спотовая цена — текущая цена на рынке, по которой актив (товар, ценная бумага, валюта) может быть куплен или продан с немедленной поставкой.

Задание 1. Анализ и визуализация данных

На первом шаге в качестве Data Scientist в компании McKinsey тебе предстоит научиться анализировать данные и графически визуализировать их изменения с течением времени.

Утром ты получил письмо от руководителя аналитического направления с инструкцией по выполнению первого задания.

Привет!

Мы столкнулись с необходимостью анализа данных для нашего клиента — крупной транснациональной компании, добывающей природный газ. Судя по их информации, Австрия, которая закупает газ у компании, не сможет обеспечить себя на 100% «зеленой» энергией, производимой с помощью ветровых установок и солнечных панелей. Но прежде чем использовать реальные данные, которые компания планирует прислать нам в ближайшее время, мы просим тебя проанализировать открытый дата-сет **opsd_austria_daily.csv**.

Наш Senior Data Scientist Олег составил для тебя шаблон по визуализации исходных данных в виде гистограмм и распределений во времени (McKinsey_task1_template.py, который доступен во вкладке «Полезные материалы»).

Твоя задача состоит в том, чтобы **правильно подтянуть данные из исходного файла опираясь на комментарии, оставленные Олегом**. Ты также должен **дополнить код для того, чтобы изучить сезонность потребления электроэнергии, а также производства солнечной и ветровой энергии**. Для этого тебе нужно сфокусироваться на одном годовом цикле, например на 2019 годе, и построить графики для Electricity_consumption, Wind_production и Solar_production за этот период. Есть несколько способов определения трендов во временных рядах, один из которых связан с использованием скользящего среднего. Это когда для каждой точки временного ряда берется среднее значение точек по обе стороны от нее, что позволяет сглаживать шум и влияние сезонности, а также выделить определенный тренд. **Не забудь оставить комментарий в коде о том, как меняется производство ветровой и солнечной энергии, а также потребление электричества в течение года (1-2 предложения)**.

Hints. В Pandas можно выбрать конкретный период с помощью запроса df.loc['start-date':'end-date']['Electricity_consumption'], который нужно написать внутри функции plt.plot(). Для rolling average используй window=15.

Ждем обновленный шаблон. Спасибо!

Полезные материалы

Шаблон кода, который подготовил Олег.

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

Задание 2. Использование Decision Tree + AdaBoost для прогнозирования временных рядов

После успешного завершения анализа и визуализации данных тебе нужно построить достоверный прогноз временных рядов с использованием несложного, но эффективного алгоритма на основе Machine Learning[3].

На почте ты обнаружил новое письмо от коллег из направления Analytics с постановкой очередного задания.

Добрый день!

Поздравляем, ты хорошо справился с задачей по визуализации и анализу сезонности исходных данных.

Теперь нам нужно спрогнозировать с помощью ML-алгоритма динамику временного ряда Electricity_consumption. Олег посоветовал начать с простых методов, например ансамблевого обучения (ensemble learning)[4]. Для начала можно попробовать объединить метод на основе деревьев принятия решений с алгоритмом AdaBoost.

Наш Senior Data Scientist помог тебе с написанием кода по применению метода Decision Tree + AdaBoost для прогнозирования значений потребления электричества в 2019 году (McKinsey_task2_template.py, который доступен во вкладке «Полезные материалы»).

Твоя задача состоит в том, чтобы **заполнить файл исходя из комментариев, оставленных Олегом, в частности применить метод fillna для заполнения пропущенных значений**. Ты также должен **исследовать влияние гиперпараметров max_depth (максимальная глубина деревьев) и n_estimators (число деревьев) на точность модели**. Из предложенных значений для двух параметров тебе нужно выбрать такие, при которых Root Mean Square Error (RMSE) и Mean Absolute Error (MAE) минимальны. Помимо способа, предложенного в шаблоне, ты можешь использовать GridSearch для подбора параметров модели.

Как обычно, ждем обновленный шаблон. Спасибо!

Полезные материалы

Шаблон файла с кодом, который подготовил Олег.

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

[3] Machine Learning (ML, машинное обучение) — методики анализа данных, которые позволяют аналитической системе обучаться в ходе решения множества сходных задач.
[4] Ансамблевое обучение использует несколько алгоритмов обучения, чтобы повысить точность модели, сделать ее более надежной и стабильной.

Задание 3. Разработка бизнес-гипотез на основе анализа данных и оценка их экономического эффекта

Твое погружение в исследование данных проходит очень успешно, впереди осталось только финальное задание.

Тем временем на почте появилось новое письмо с инструкцией, что делать дальше.

Привет!

В качестве последнего задания в роли Data Scientist в McKinsey тебе предстоит выдвинуть несколько бизнес-гипотез на основе проведенного анализа данных.

В частности, мы предлагаем тебе **вычислить разность между потреблением электричества и производством энергии с помощью ветровых и солнечных установок** и добавить эту величину в исходный файл **opsd_austria_daily.csv** последним столбцом Delta и загрузить решение, переименовав файл в **opsd_austria_daily_updated.csv**.

Нахождение разницы между потреблением и производством энергии важно для расчета объема ресурсов в энергетическом портфолио страны. Таким образом, ты **сможешь доказать или опровергнуть гипотезу клиента** о том, что Австрии не хватит собственных ресурсов ветряной и солнечной энергии для компенсации спроса на электроэнергию, которую страна вынуждена будет покрывать за счет импорта природного газа. Напиши в файле кода несколько строк комментариев по этому поводу.

Наконец, тебе необходимо определить примерную стоимость Delta, используя спотовую цену и указать значения в столбце Price_delta, который также нужно добавить в файл **opsd_austria_daily_updated.csv**. Это позволит компании оценить приблизительную прибыль в случае, если Австрия будет закупать у них газ.

Hints. Не забудь, что спотовая цена указана для 1 кВт·ч, а энергия дана в ГВт·ч.

Существует несколько методов добавления столбцов в pandas Dataframe. Можешь ознакомиться с ними в статье, приведенной во вкладке «Полезные материалы».

Чтобы сохранить обновленный файл в формате .csv, используй команду df.to_csv('opsd_austria_updated.csv').

Успехов!

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы, включая файл с кодом .py и обновленный .csv файл.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.