Отчет о научно-исследовательской работе

Список исполнителей

- Антон Агейков - Data Scientist, капитан команды
- Асем Ибраева - Data Scientist
- Тимофей Акимкин - ML Engineer
- Яна Бучковски - технический писатель
- Татьяна Плевако - DevOps

Куратор: Артем Карасюк

Реферат

Пандемия продолжает влиять на страны по всему миру - COVID-19 затронул 195 стран с примерно 366 млн подтвержденными случаями заболевания (к январю 2022). Понимание динамики передачи инфекции в каждой стране и прогнозы имеют решающее значение для дальнейших действий по борьбе с пандемией. Целью проекта является разработка и визуализация модели, которая предсказывает заболеваемость COVID-19.

Содержание

1. Термины и определения

- **Датасет** - размеченный набор данных.
- **Дэшборд (инфопанель)** - это аналитическая панель с понятным интерфейсом для интерактивного взаимодействия с огромным количеством постоянно изменяющихся показателей.
- **Отставание** - предполагаемое количество дней до наступления волны заболеваемости COVID-19 в выбранной стране.
- **Последовательность работ (pipeline)** - это последовательность стадий, внутри которых расположены задачи. Расположены они таким образом, что выход каждого элемента является входом следующего.
- **Реляционная база** - это набор данных с предопределенными связями между ними. Эти данные организованы в виде набора таблиц, состоящих из столбцов и строк.
- **Степень похожести (сходства) стран** - коэффициент корреляции Пирсона двух временных рядов, соответствующих числу заболевших за день в каждой из стран.
- **MAE, Mean Absolute Error (средняя абсолютная ошибка)** - среднее арифметическое модуля отклонения предсказанного значения от реального.
- **MAPE, Mean Absolute Percent Error (средняя абсолютная ошибка в процентах)** — эта метрика показывает, на сколько процентов в среднем предсказанное значение отклоняется от реального значения.
- **MySQL** — свободная реляционная система управления базами данных.
- **RMSE (Root Mean Square Error, RMS Error)** - среднеквадратичная ошибка, расстояние между двумя точками.

2. Обозначения и сокращения

- **COVID-19** - коронавирусная инфекция 2019 года, вызываемая коронавирусом SARS-CoV-2.
- **CI/CD** (от англ. Continuous Integration, Continuous Delivery — дословно «извлечение, преобразование, загрузка») - это технология автоматизации тестирования и доставки новых модулей разрабатываемого проекта заинтересованным сторонам.
- **ETL** (от англ. Extract, Transform, Load — дословно «извлечение, преобразование, загрузка») – это процесс извлечения, преобразования и загрузки данных.

3. Введение

Первым делом была сделана гипотеза о том, что сценарий распространения заболевания в разных странах имеет сходство. Было решено численно

сравнить максимальную схожесть стран при отставании в несколько дней. В планах было рассчитать степень схожести двух стран, исходя из графиков ежедневного прироста заболеваемости в каждой стране.

Степень схожести определялась как пирсоновская корреляция двух дискретных функциональных зависимостей прироста количества заболевших в каждой стране. Отобразив полученные степени схожести стран на географической карте, рассчитывалось найти страны с наиболее похожим сценарием на сценарий распространения COVID-19 в России. Согласно сделанной гипотезе, сценарии развития заболеваемости стран с наибольшей степенью схожести с Россией будут с большой степенью вероятности схожи со сценарием в России, что позволит прогнозировать заболеваемость.

Затем было сделано еще несколько предположений насчет использования других моделей прогнозирования и сравнения их метрик для выбора наиболее точной. Подробная информация о моделях и их метриках будет представлена в разделе **4.4 Рассмотренные способы прогнозирования**.

4. Основная часть

Для работы над проектом были выбраны *ETL* и *Yandex DataLens*. Работа была разделена на три части:

- **обработка данных, ETL** – процесс извлечения, преобразования и загрузки данных;
- **CI/CD** - доставку кода на сервер;
- **визуализация** - создание дэшборда в *Yandex DataLens* с возможностью выбрать страну, визуализировать степень схожести страны с другими, прирост заболеваемости по дням в виде графика.

4.1. Обработка данных

Данные о заболеваемости и смертности в разных странах были взяты из открытых источников (COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University и Data on COVID-19 (coronavirus) by Our World in Data), обработаны (выделены нужные поля для обработки и визуализации), а затем по этим данным были найдены отставание и степень схожести для каждой пары стран.

Как уже было сказано ранее, степень схожести стран была определена как пкоэффициент корреляции Пирсона двух временных рядов, соответствующих числу заболевших за день в каждой из стран. Данные выгружались в реляционную базу *MySQL* и S3-хранилище на *Yandex Cloud*.

4.2. CI/CD

При коммите в репозиторий: 1) Обновляется конфигурация для запуска пайплайна:

```
1.1) Обновление секретов.

1.2) Обновление пакетов *python*.
```

2) Запускается пайплайн, который определён в `dodo.py`:

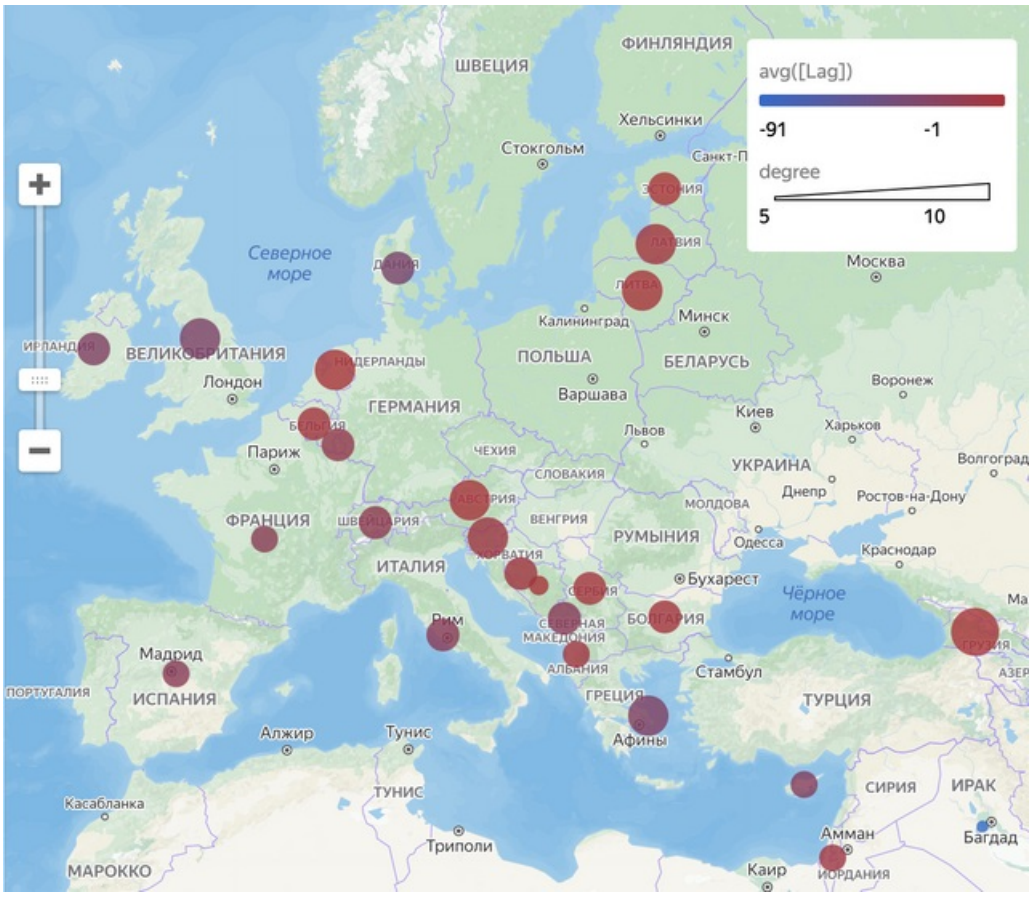
```
2.1) Подкачивает новые данные.

2.2) Предобрабатывает их для последующей обработки.

2.3) Высчитывается похожесть стран.

2.4) Высчитывается предсказание моделей.

2.5) Загружает все данные в S3.

2.6) Загружает данные для визуализации в *MySQL*.
```

4.3. Визуализация

Dashboard

Визуализация предполагала отображение на географической карте степени схожести страны с выбранной пользователем из списка и отставание сценария согласно модели.

Создание географической карты с данными происходило таким образом: на сервисе *Yandex DataLens* создавалось новое подключение к таблице данных из реляционной базы *MySQL*. Далее добавлялся новый датасет на основе таблицы данных, содержащей название страны, степень схожести, отставание, широту и долготу точки на территории страны. Создавалось поле географической координаты при помощи функции `GEOPOINT` от двух аргументов - широты и долготы точки. Затем на карте были проставлены точки.



Размер точки отображает степень схожести страны, а цвет - количество дней отставания. В качестве цветовой гаммы был выбран непрерывный градиент, цвет каждой точки автоматически определялся сервисом в зависимости от отставания.

Используемые переменные для отображения карты схожести стран:

- `longitude` (Долгота) - поле, содержащее долготу одной точки страны.
- `latitude` (Широта) - поле, содержащее широту одной точки страны.
- `Coord` (Координаты точки) - поле, содержащее координаты одной точки страны.
- `Страна для отображения` - поле, содержащее название страны. Используется для выбора по селектору (в качестве фильтра).
- `Другая страна` - поле, содержащее название не выбранных стран для отображения. С ними происходит сравнение.
- `Степень отставания` - поле, содержащее количество дней отставания для двух выбранных стран (см. Термины).
- `Степень уверенности` - поле, содержащее степень сходства двух выбранных стран (см. Термины).

Также были созданы графики новых случаев госпитализации пациентов, больных COVID-19 в выбранной стране на протяжении выбранного периода времени, а также график смертности, заболеваемости и вакцинации.

Ниже можно увидеть график смертности, заболеваемости и вакцинации.

Используемые переменные для отображения графиков смертности, заболеваемости, вакцинации и госпитализаций:

- `location` (Местоположение) - поле, содержащее название страны, для которой отрисовывается график случаев заражений, вакцинаций, смертей.
- `date` (Дата) - поле, содержащее дату в формате "ГГГГ-ММ-ДД".

- `total_cases` (Все зафиксированные случаи заражения) - количество всех зафиксированных случаев заражения в определенной стране на определенную дату (с начала пандемии).
- `new_cases` (Новые случаи заражения) - количество новых зафиксированных случаев заражения в заданной стране в течение дня.
- `new_cases (smoothed)` (Новые случаи заражения (сглажено)) - те же `Новые случаи заражения`, но сглаженные при помощи фильтра высоких частот.
- `total_deaths` (Все смерти) - поле, содержащее общее количество зафиксированных случаев смерти от коронавируса в определенной стране с начала пандемии на определенную дату.
- `new_deaths` (Новые случаи смерти) - количество новых зафиксированных случаев смерти от коронавируса в заданной стране в течение дня.
- `Новые случаи смерти (сглажено)` - те же `Новые случаи смерти`, но сглаженные при помощи фильтра высоких частот.
- `Новые случаи вакцинации (сглажено)` - поле, содержащее количество вакцинированных в заданной стране в течение дня, сглаженные при помощи фильтра высоких частот.
- `Госпитализировано` - поле, содержащее количество случаев госпитализаций на определенную дату в определенной стране. Информация по госпитализациям была найдена в открытом доступе только по США.

4.4. Рассмотренные способы прогнозирования

Метрики модели схожести стран:

```
`MAE` (30 дней): 36093.759
`MAPE` (30 дней): 0.489 ≈ 48.89%
```

После рассчетов метрик основной модели (модели схожести стран), стало понятно, что данная модель дает большую ошибку относительно других моделей.

Кроме модели, составленной на основе схожести стран, были исследованы другие способы прогнозирования для сравнения полученных предсказаний. В качестве метрик были выбраны `MAPE` и `MAE` (см. Термины). И `MAPE`, и `MAE` более устойчивы (в сравнении с `RMSE`) к влиянию выбросов благодаря использованию абсолютного значения.

- **Линейная регрессия**

Как была построена линейная регрессия: в качестве признаков берутся последние 30 дней текущей страны и последние 30 дней со сдвигом на величину отставания для трёх самых похожих стран. Линейной регрессией предсказывается число, которое предсказывает заболеваемость на следующий день. Для предсказание следующего дня предсказанное принимается за истинное и окно признаков для всех стран сдвигается на один день. Так происходит 30 раз.

Метрики построенной линейной регрессии (по России):

```
`MAPE` (30 дней): 0.420 ≈ 42%
Дополнительно:
`MAPE` (60 дней): 0.386 ≈ 38.6%
```

- **Модель Prophet**

Как была построена модель Prophet: в качестве регрессоров берётся история заболеваемости для текущей страны и двух самых похожих стран со сдвигом на количество дней отставания. Все аналогично как и с линейной регрессией, но Prophet для предсказания требуется история заболеваемости из других стран, которая была получена ранее (так как были взяты страны, которые отстают от нашей больше, чем на количество дней, которые необходимо предсказывать).

Метрики модели Prophet:

```
`MAPE`(30 дней): 0.2 ≈ 20%
```

- **Модель Хольта-Винтерса**

В качестве последнего метода прогнозирования было решено рассмотреть временной ряд - «случайный процесс, состоящий из сигнала, отражающего реальную эпидемическую ситуацию, и высокочастотного шума». Экспоненциальное сглаживание может справиться с такой изменчивостью внутри ряда, сглаживая шум.

В представленных данных временные ряды имеют отчетливый тренд с сезонностью. Исходя из этого, из всех моделей фильтрации была выбрана модель Хольта-Винтерса. Она представляет собой комбинацию трех других более простых компонентов, каждый из которых является методом сглаживания:

1) *При простом экспоненциальном сглаживании (SES)* модельное значение представляет собой средневзвешенную между текущим истинным и предыдущим модельным значениями.

2) *Экспоненциальное сглаживание Хольта (HES)* позволяет нам учитывать тренд во временном ряде, при данном методе предполагается, что будущее направление изменения ряда зависит от взвешенных предыдущих изменений.

3) *Экспоненциальное сглаживание Винтерса (WES, или метод Хольта-Винтерса.)* - это расширение экспоненциального сглаживания Хольта, которое, наконец, позволяет учесть сезонность.

В параметрах модели, исходя из данных, была указана `Multiplicative trend and additive seasonality`. Мультипликативность тренда означает, что тренд нелинейный (изогнутая линия), а аддитивная сезонность означает, что ширина или высота сезонных периодов не меняется с течением времени.

Метрики модели Хольта-Винтерса:

```
`MAE` (30 дней): 281972.607

`MAPE` (30 дней): 0.024 ≈ 2.4%
```

Как видно из результатов, модель Хольта-Винтерса работает лучше при краткосрочных прогнозах (в данном случае 30 дней), так как при экспоненциальном сглаживании прошлые наблюдения взвешиваются в экспоненциально убывающем порядке. Это означает, что самым последним наблюдениям присваивается более высокий вес, чем далеким значениям. В случае России заболеваемость на последнем временном отрезке пошла резко вверх, что и сказалось на прогнозе до начала марта. **Данная модель была в итоге выбрана основной, так как оказалась наиболее точной.**

5. Заключение

Гипотеза о том, что сценарий распространения заболевания в разных странах имеет сходство, работает с большой ошибкой, так как модель, основанная на данном предположении, имеет `MAPE` больше 20%. Также было выполнено сравнение прогнозов с прогнозами, полученными другими моделями, например, с моделью Prophet и получены удовлетворительные результаты точности. **Модель Хольта-Винтерса** показала лучшие результаты (`MAPE` ≈ 2.4%) и была выбрана в итоге основной. На основе рассчитанных данных были созданы прогнозы на 30 дней для России и ряда других стран.

6. Список источников

1. Зараза, гостья наша. Как математика помогает бороться с эпидемиями.
2. Коронавирус: анализ данных без паникерских настроений.
3. О. И. Криворотько, С. И. Кабанихин. Математические модели распространения COVID-19 [Текст] / О. И. Криворотько С. И. Кабанихин – Новосибирск, 2022. – 64 с.
4. Пандемия COVID-19 глазами математика, или почему классическая модель SEIRD не работает.
5. COVID-19: Модель параметрического предсказания эпидемии.
6. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.
7. Data on COVID-19 (coronavirus) by Our World in Data.