Для предсказания распространения использовались данные о

1. **Истории распространения коронавируса:** ежедневное кол-во заболевших.
2. **Социально-экономическом состоянии**: плотность и количество населения, доля рабочего населения, данные о качестве жизни, количество приехавших людей в общем, количество людей, приехавших из Москвы (данные от tutu), количество туристов за предыдущие года, доля городского и сельского населения в разном возрасте.
3. **Качестве здравоохранения**: данные о больных инфекционными заболеваниями за март, epirank индексы, количество больных туберкулезом за предыдущие года, количество аппаратов ИВЛ и ЭКМО на 100 тысяч человек.
4. **Самоизоляции**: индекс самоизоляции от Яндекса.

Для прогнозирования был выбран алгоритм регрессии, в частности lightgbm. Было использовано две модели: одна для предсказания количества новых зараженных на следующий день, вторая - для предсказания индекса самоизоляции на следующий день. После предсказания для нужного периода использовалась кумулятивная сумма для формирования представления общего количества.

**Каким образом строились признаки для прогнозирования t + 1 дня в определенной стране**:

Данные из пункта 2-3 напрямую вставлялись в вектор признаков.

На основании данных из пункта 4 добавлялись:

1. Предсказание коэффициента самоизоляции экспоненциальным сглаживанием на следующий день
2. Лаги 1-14
3. Скользящие среднее, стандартное отклонение, максимальное, минимальное.
4. Кол-во населения разделить на коэффициенты самоизоляции за последние 6 дней (предполагая, что 6 дней - это инкубационный период и индекс самоизоляции обратно пропорционален кол-ву зараженных)

На основании данных истории распространения коронавируса строились следующие признаки:

1. Лаги 1-14 по количеству новых зараженных
2. Скользящие среднее, стандартное отклонение, максимальное, минимальное количества зараженных.
3. Максимальные значения количества зараженных за все время до предсказываемого дня (пик)
4. Лаги 1-3 минус пик зараженных (из предыдущего пункта)
5. На лагах за последние 7 дней строится ridge регрессия, и в признаки добавляются значения коэффициентов и предсказание на 8ой день. Решающие деревья не могут ловить тренды, и эти признаки предназначены помочь им с этим
6. Строится полином, аппроксимирующий новые заражения за последние 7 дней, в признаки добавляются коэффициенты этого полинома и предсказание для 8ого дня. Эти признаки дают модели информацию, является ли рост случаев экспоненциальным

Так как предсказания нужно делать больше чем на 1 день вперед, модель работает рекурсивно.

1. На основании имеющихся данных за t дней делаются предсказания для t + 1 дня
2. С учетом предсказания на t + 1 день считаются признаки, использующие данные о прошлом, для t + 2 дня
3. Делается предсказание для t + 2 дня

Таким образом делаются предсказания для всех дней до конца года.

В качестве гиперпараметров lightgbm использовались те, которые давали лучшие результаты при валидации на последней неделе доступных данных.