Для предсказания распространения использовались данные о

1. **Истории распространения коронавируса:** ежедневное кол-во заболевших, выздоровевших, умерших.

Очевидно, что для прогнозирования будущего необходимо знать прошлое.

1. **Социально-экономическом состоянии страны**: ВВП, количество мигрантов, процент городского населения, плотность населения, общий объем населения, доля торговли в ввп, количество вылетов/прилетов за 2018 год, геоположение страны.

Жители развитых стран более мобильны и чаще взаимодействуют с жителями других стран, что способствует распространению вируса

1. **Качестве здравоохранения**: кол-во смертей от рака, туберкулеза, детская смертность, средняя продолжительность жизни, процент смерти от неинфекционных заболеваний, количество больничных коек на 1000 человек, процент людей доживших до 65 лет, коэффициент выявления случаев туберкулеза, коэффициент смертности среди взрослого населения, количество курящих

Эти данные напрямую связаны с вирусом. В странах с развитой системой здравоохранения вирусу сложнее распространяться, и такие признаки позволяют модели уловить эти зависимости.

1. **Принятых мерах**: был ли введен карантин, на каком уровне был введен карантин.

Аналогично пункту 3, принятые напрямую влияют на ситуацию с болезнью внутри страны. Эти данные помогают модели оценить, как карантин влияет на распространение вируса

Для прогнозирования был выбран алгоритм регрессии, в частности lightgbm. Было использовано три таких модели: для предсказания количества подтвержденных случаев, умерших и выздоровевших. Основной задачей каждой модели было по данным за t дней предсказать количество новых случаев (заражения/выздоровления/смерти в зависимости от модели) за t + 1 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. Количество дней с начала карантина. Так как карантин начинает действовать спустя какое-то время, модель пытается уловить эту зависимость.
2. Категориальный признак - тип карантина. Карантин на уровне страны, ожидаемо, более весомый чем на уровне города.

На основании данных истории распространения коронавируса строились следующие признаки:

1. Лаги за последние семь дней по каждой целевой переменной ( кол-во случаев заражения/выздоровления/смерти ) за последние 7 дней (t-7;t]
2. Среднее, минимальное и максимальное значения целевых переменных и стандартное отклонение их распределений за предыдущую неделю (t-7;t]. Эти признаки дают модели информацию о распределении целевых переменных в последнее время.
3. Максимальные значения целевых переменных за все время до t + 1 дня (-inf; t]. Этот признак дает некое представление о пике случаев на t день.
4. Значения числа зараженных за предыдущие 3 дня минус максимальное кол-во зараженных в день за все время до данного дня ( признак из пункта 3 ). Этот признак дает модели понять, как пиковое значение отличается от последних данных по заражению.
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 целевые переменные 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.