Описание кейса

Вирус COVID-19 начал распространяться внутри России в 2020 году. Для наиболее полного анализа распространения вируса правильно рассмотреть случаи заражения по каждому отдельному городу. Набор данных включает информацию о случаях заражения в 842 средних и крупных городах России за март 2020 года. Необходимо предсказать уровень заражения вирусом в городах, считая метрикой качества построенной модели MAE (mean absolute error).

Период рассмотрения: 01/03/2020–30/03/2020

Объекты: 842 города в 84 регионах России

Целевая переменная:

* *Inf rate* — частота заражения, рассчитывается как Log(infectedat30.03 + 1) - Log( infectedat\_15.03 + 1)

Зависимые переменные

* *whole\_population* — население по каждому населенному пункту
* *urban/Rural* – количество городских и сельских жителей
* *name* – название населенного пункта
* *district* – федеральный округ
* *region\_x* – регион
* *density* – плотность населения
* *Lat* – северная широта населенного пункта
* *Lng* – восточная долгота населенного пункта
* *cleanness, publicservices, neighbourhood, childrenplaces, sportandoutdoor, shopsandmalls, publictransport, security, lifecosts* – рейтинг городов по качеству жизни, данные Domfond.ru
* ivlper100k, ivlnumber, ekmoper100k, ekmonumber — количество аппаратов искусственной вентиляции легких в абсолютном выражении, на 100 тыс. человек населения, количество оборудования для ЭКМО — в абсолютном выражении, на 100 тыс. человек населения.
* *infected*3003, died*3003, recovered*3003, sick*3003, infected*1503, died*1503, recovered*1503, sick*1503* – мера для расчета уровня заражения в регионе. Для анализа мы считаем, что инкубационный период равен двум неделям, и вычисляем логарифмически преобразованный прирост как показатель скорости заражения.
* *avgtempmin, avgtempmax, avgtempstd, avgtempmedian,  
  humiditymin, humiditymax, humiditystd, humiditymedian,  
  pressuremin, pressuremax, pressurestd, pressuremedian,  
  windspeedmsmin, windspeedmsmax, windspeedmsstd, windspeedmsmedian* – данные погодных условий за март 2020 года
* *urban50–54years, urban55–59years, urban60–64years,  
  urban65–69years, urban70–74years, urban75–79years,  
  urban80–84years, urban85–89years, urban90–94years,  
  rural50–54years, rural55–59years, rural60–64years,  
  rural65–69years, rural70–74years, rural75–79years,  
  rural80–84years, rural85–89years, rural90–94years* – количество жителей по возрастным группам и районам проживания (город/село)
* *workratio15–72years, workratio55–64years, workratio15–24years, workratio15–64years, workratio25–54years*– занятность населения по возрастным группам
* *nump patientstuberculosis1992 .. 2017* — количество больных туберкулезом в населенных пунктах по годам
* *volumeservhousehold2017, volumeservchargeable2017, volumeservtransport2017, volumeservpost2017, volumeservaccommodation2017, volumeservtelecom2017, volumeservothers2017, volumeserveterinary2017, volumeservhousing2017, volumeserveducation2017, volumeservmedicine2017, volumeservdisabled2017, volumeservculture2017, volumeservsport2017, volumeservhotels2017, volumeservtourism2017, volumeservsanatorium2017* — объем предлагаемых населению услуг в рублях
* *numphonesrural2018, numphonesurban2018*— количество телефонов в разбивке по городским и сельским районам
* *has\_metro* – наличие метро в городе
* *busmarchtravel18, busapriltravel18* —пассажирооборот автобусов по маршрутам регулярных перевозок (тысяча пассажиро-километров)
* *epirankavia, epirankbus, epiranktrain, epirankaviacat, epirankbuscat, epiranktraincat* — рейтинг экологической безопасности, индекс epirank