Резюме.

В процессе выполнения задания были загружены и предобработаны предложенные данные по заболеваемости новой короновирусной инфекцией.

Изучена теория и отработана практика анализа и прогнозирования временных рядов с помощью языка программирования R.

Данные делились на обучающую и тестовую выборки по дате, отсекающей для тестовой выборки 7 месяцев от момента окончания наблюдений (1.06.2022).

С помощью моделей экспоненциального сглаживания (ETS), тета-модели, интегрированной модели авторегрессии скользящего среднего (ARIMA) в нескольких модификациях, а также с помощью моделей машинного обучения (линейная регрессия, случайный лес, градиентный бустинг, метод опорных векторов) была предсказана заболеваемость новой короновирусной инфекцией в Москве на период с 1.06.2022 до 1.01.2023 г. Модели оценивались по среднеквадратичной ошибке (RMSE).

Наилучшие предсказания получены с помощью моделей случайного леса и метода опорных векторов.

Код доступен в публичном репозитории на GitHab по ссылке: https://github.com/alymova-yu/ts_covid

Ход работы и обсуждение.

Для работы использовался язык программирования R 4.3.1. с дополнительными библиотеками, представленными в коде.

Из предоставленного источника загружены данные по заболеваемости новой короновирусной инфекцией в регионах России в период с 12.03.2020 г по 13.01.2023 г. Также из источника загружены данные о количестве тематических запросов в сети Интернет. Список запросов представлен в таблице 1.

Таблица 1. Перечень тематических запросов.

"антитела"	"вторая волна"	"вызвать скорую"
"доставка еды на дом"	"как не заразиться"	"купить антисептик"
"купить маску и респиратор"	"лечение коронавируса"	"пропало обоняние"
"пульсоксиметр и сатурация"	"сдать тест"	"сделать КТ"
"симптомы коронавируса"	"что делать дома"	"что делать если не едет скорая"

Данные о заболеваемости отфильтрованы по региону – выделена выборка по Москве. Графическое представление динамики заболеваемости, а также изображение автокорреляционной функции представлено на рисунке 1.

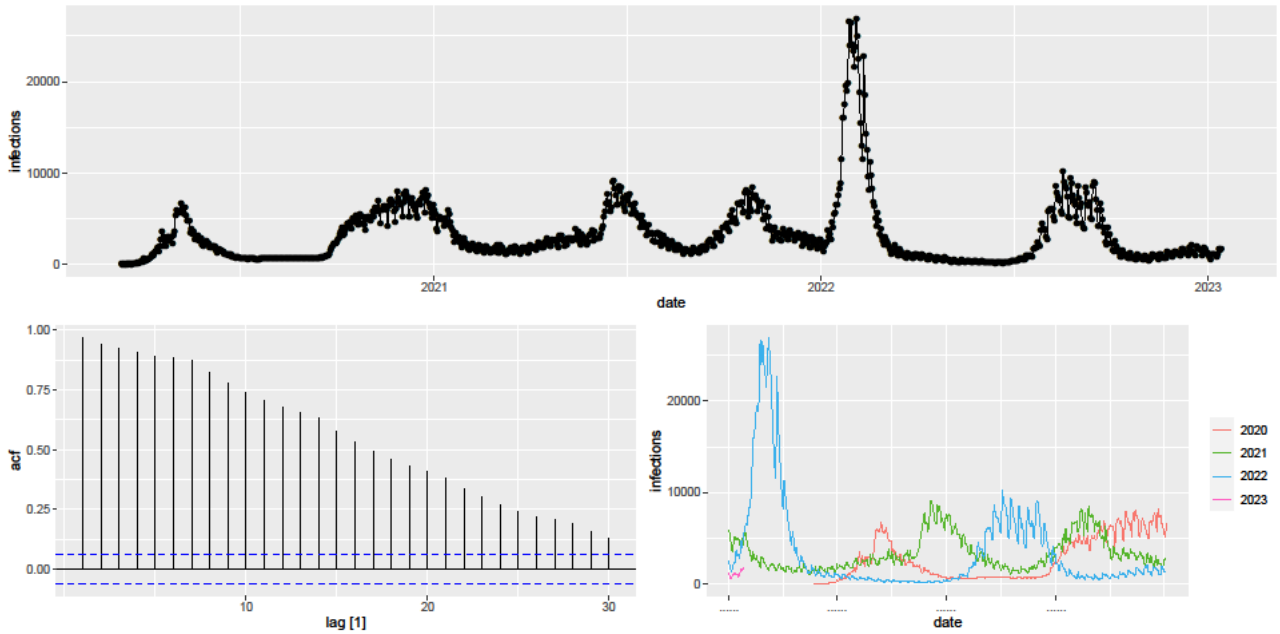

Рисунок 1. Динамика заболеваемости в Москве. Верхний график – динамика заболеваемости, нижний слева – автокорреляционная функция (acf)

На рисунке 1 видно, что выраженной внутригодовой сезонности в ряду нет. Регулярно повторяются пики подъема заболеваемости – с частотой 1 раз в 5-7 месяцев. По автокорреляционной функции можно заподозрить, что в данных может присутствовать недельная сезонность (значения acf чуть выше соседних на +7 и +14 дни).

Для оценки тренда и сезонности выполнено STL-разложение (рис. 2). STL позволяет подтвердить наличие недельной сезонности в изучаемых данных. Данный феномен может быт объясним источником данных – положительными результатами ПЦР-тестирования, результаты которых становились доступны в большем числе в рабочие дни и в меньшем – в выходные (в соответствии со стандартным режимом работы учреждений здравоохранения).

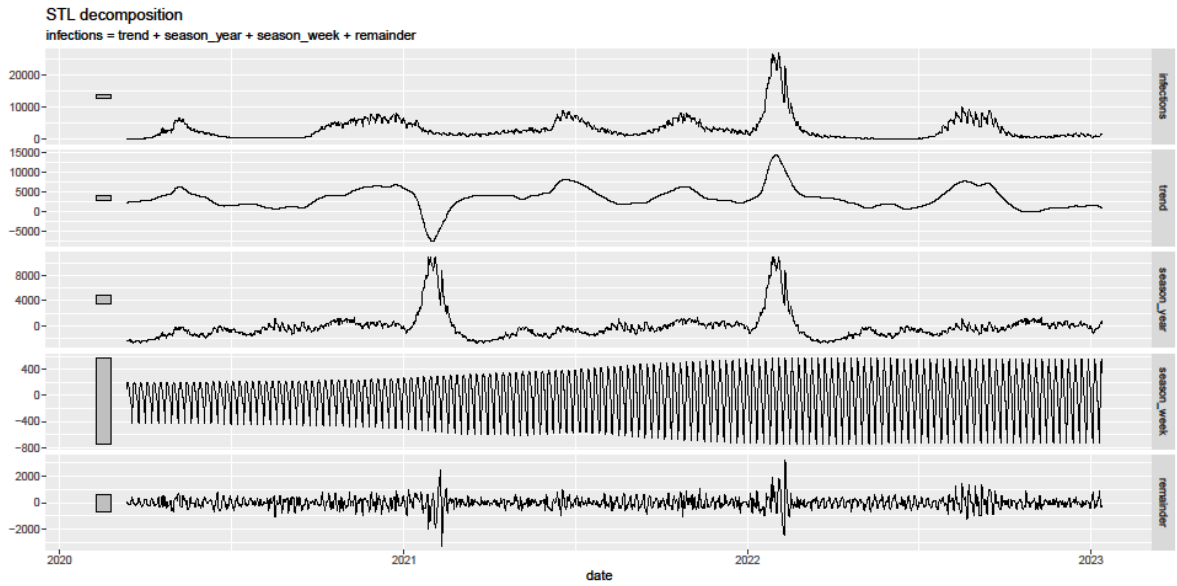

Рисунок 2. STL-разложение.

Данные по заболеваемости были разделены на тренировочную и тестовую выборки по дате, отделяющей последние 7 месяцев наблюдений (01.06.2022 г). Данные до этой даты отнесены к тренировочной выборке, после – к тестовой.

На тренировочных данных обучены модели прогнозирования временных рядов: наивная модель, модель экспоненциального сглаживания с автоподбором типа модели по критерию Акаике, тета-модель, ARIMA модель с автоподбором параметров и ARIMA модель с параметрами сезонности, полученными преобразованием Фурье.

На тестовой выборке были построены прогнозы и оценена их точность по метрике RMSE. Результаты представлены в таблице 2.

Так, лучшие результаты показали модели ETS (A,Ad,A) и ARIMA с преобразованием Фурье. Данные результаты объясняются учетом в этих моделях недельной сезонности, что можно увидеть на графическом отображении их предсказаний (рис.3).

Таблица 2. Результаты оценки качества моделей по метрике RMSE

Модель	RMSE
Наивная	3705
Тета	3588
ETS(A,Ad,A)	3151
ARIMA	3743
ARIMA фурье	3519

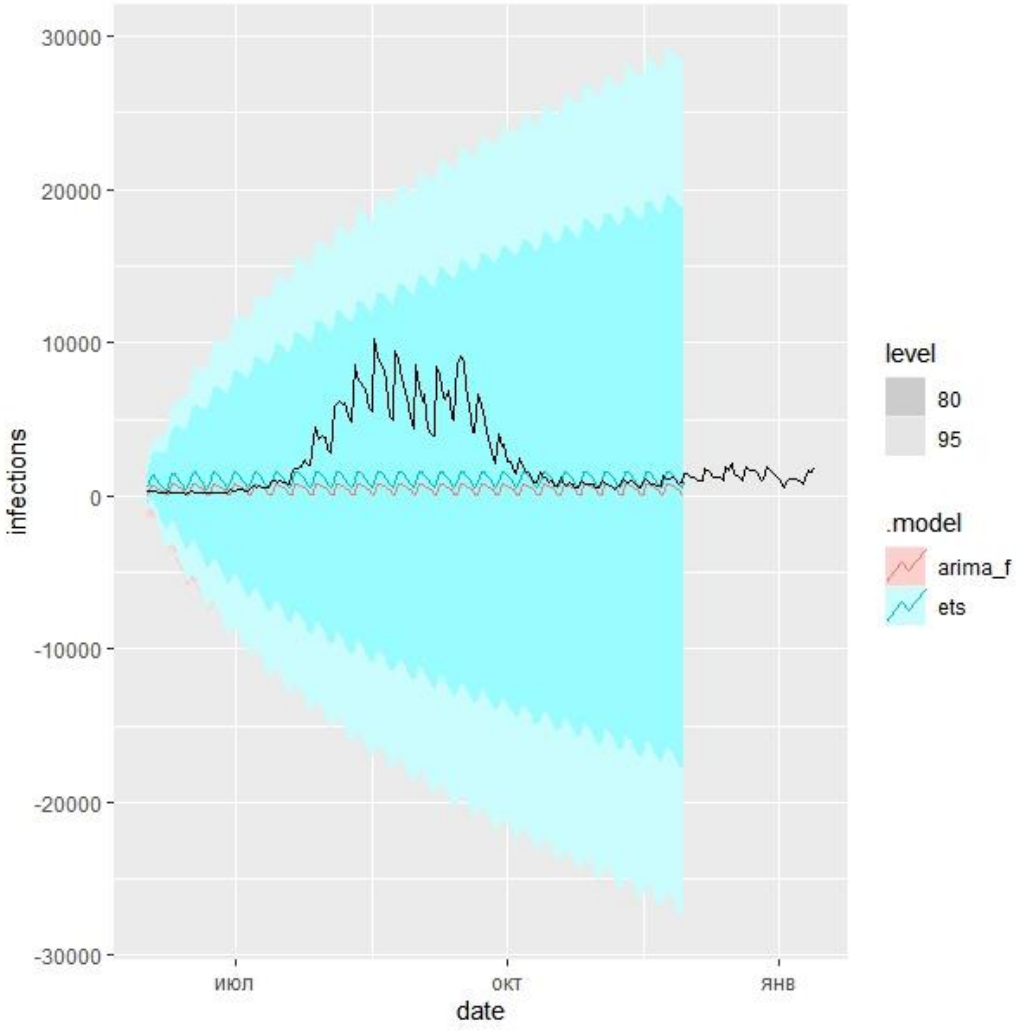

Рисунок 3. Предсказания моделей с наилучшими оценками качества. Черная линия – тестовые данные.

Для улучшения качества прогнозов решено попытаться спрогнозировать заболеваемость с помощью моделей машинного обучения.

В отличие от классических моделей, модели машинного обучения нуждаются в подборе и конструировании признаков для обучения и прогнозирования. В случае с временными рядами эта задача – творческая.

В загруженных данных присутствуют как количества новых зараженных на каждый день, так и тематические запросы в сети Интернет, которые можно использовать в качестве признаков. Однако в задаче прогнозирования в прогнозируемый момент времени отсутствуют данные о значениях признаков – в отличие от стандартных задач машинного обучения, в которых по признакам объекта в данный момент времени нужно предсказать целевую переменную. В связи с этим конструирование признаков в данном случае решено сделать тремя путями:

1) Часть признаков – время, а также цикличная недельная сезонность (в виде тригонометрических признаков, полученных преобразованием Фурье) рассчитывалась для каждого момента времени
2) Добавлены признаки, несущие в себе информацию о числе зараженных в моменты времени на 7-6 месяцев раньше, чем предсказываемый момент (t-212, t-196, t-182), и разницы между этими моментами.
3) Добавлены признаки, представляющие собой разницу между количеством тематического запроса в моменты времени на 7-6 месяцев раньше момента прогноза – для всех перечисленных в таблице 1 запросов.

Таким образом, для обучения и прогнозирования стал доступен набор данных с 29 признаками и 1 целевой переменной.

Поскольку конструирование признаков в данном случае предполагает обращение к данным на 6-7 мес более ранним, для первых 7 месяцев выборки

их построение невозможно – данный период исключается из тренировочных данных.

Модели машинного обучения (линейная регрессия – LS, метод опорных векторов – SVM, случайный лес – RF, градиетный бустинг – GB) обучались на наборах данных с разными комбинациями признаков, для подбора параметров моделей использовалась кросс-валидация. Прогнозы моделей оценивались с помощью метрики RMSE. Оценки моделей с разными комбинациями признаков представлены в таблице 3.

Таблица 3. RMSE для оцениваемых моделей

Признаки	SVM	LS	RF	GB
Число заболевших	2501.559	5669.369	3048.507	3709.957
Разница в числе заболевших	2694.945	6093.759	3080.104	3318.684
Число + разница	2148.786	6019.435	2928.803	3805.248
Запросы	3374.226	3118.513	3263.291	3312.338
Число + разница + запросы	2074.648	6371.745	2640.712	3468.051

Наилучшие прогнозы показала модель SVM (что согласуется с данными литературы), минимальное значение RMSE получено для модели SVM, работающей на 29 признаках. Наличие информации о запросах также увеличило предсказательную силу и для случайного леса. Графическое отображение предсказаний для этих моделей представлено на рисунке 4.

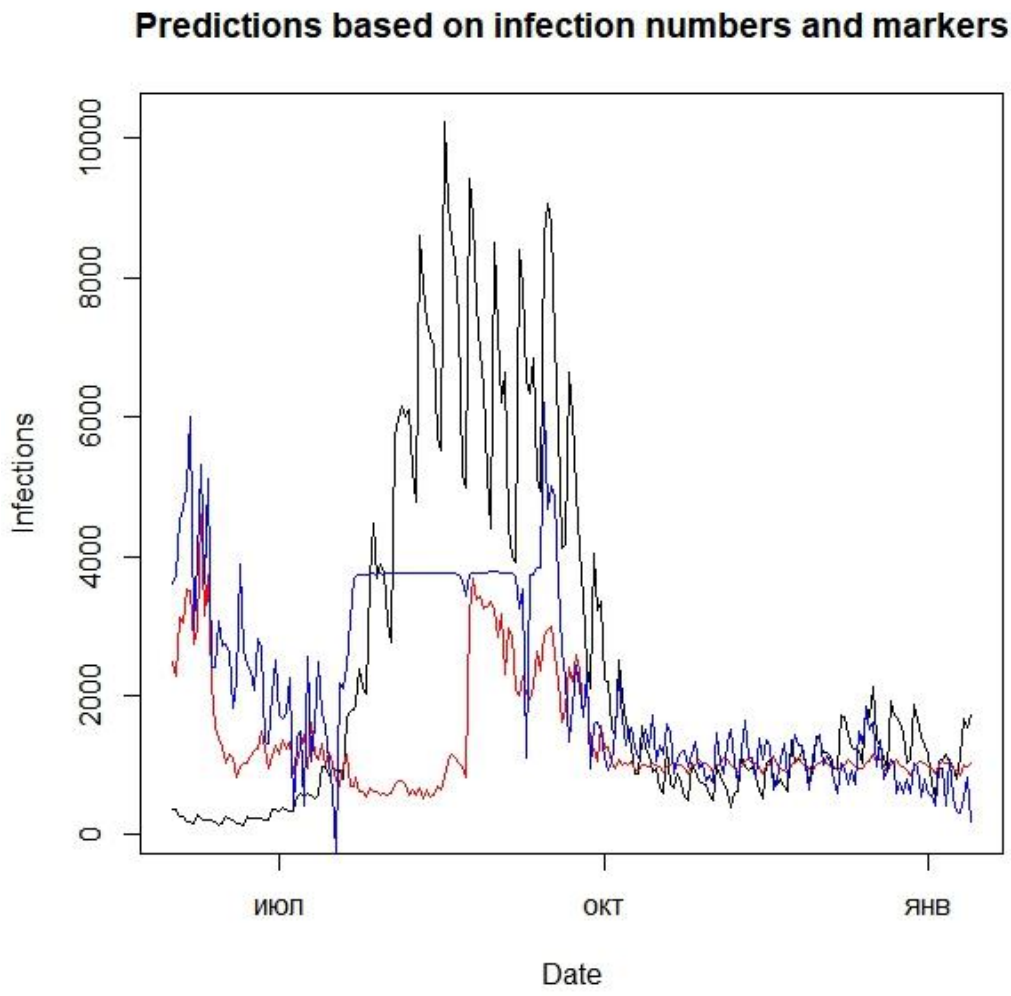

Рисунок 4. Предсказания моделей машинного обучения (SVM – синий, RF – красный), обученных на выборках с 29 признаками (включая информацию о запросах) и реальные тестовые данные (черный).

Заключение.

В данной работе проведена попытка прогнозирования заболеваемости новой короновирусной инфекцией на период 6 месяцев. С данной задачей заметно лучше справилась модель машинного обучения SVM, обученная на данных с 29 признаками (тригонометрические признаки, скользящие данные о заболеваемости в прошлом, данные о тематических запросах в прошлом).

Согласно изученным литературным источникам, надежное прогнозирование уровня заболеваемости новой короновирусной инфекцией с помощью классических моделей и моделей машинного обучения ограничивается периодом в 1-2 месяца. Это связано с большим количеством сложноучитываемых факторов, влияющих на заболеваемость (появление новых штаммов, ограничительные и профилактические меры, государственная политика в области здравоохранения и проч).

Наиболее перспективным видится использование методов глубинного обучения с учетом временных рядов заболеваемости, географических данных, а также данных о штаммах и профилактических мерах.