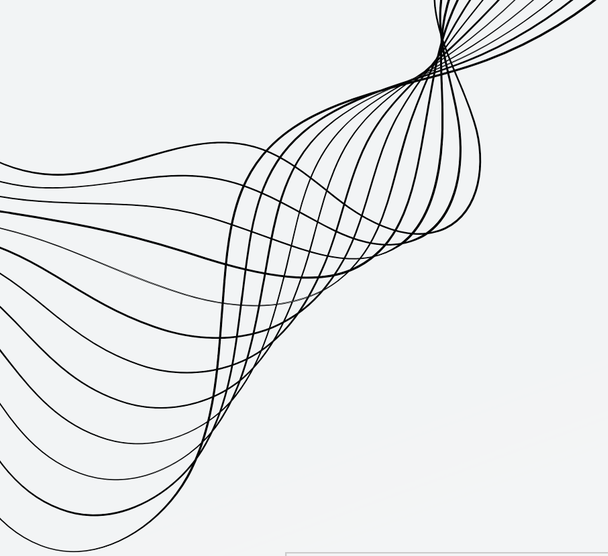


ПОЛИТЕХ
Физико-механический институт
Высшая школа теоретической механики
и математической физики

ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ РАЗВИТИЯ ПАНДЕМИИ СРЕДСТВАМИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Выполнила:
студентка гр. 5040103/10301 М.А. Курдина

Руководители:
доцент, к.ф.-м.н. А.А. Ле-Захаров
ассистент Д.С. Перец

СОДЕРЖАНИЕ

СТАТИСТИКА COVID-19 В США

Случаи

Новые случаи (всего за неделю)

414 721

Nov 2022 Jan 2023

Летальные исходы

Новые смерти (всего за неделю)

3 907

Nov 2022 Jan 2023

Госпитализации

Новые поступления (ежедневное среднее)

5 630

Nov 2022 Jan 2023

Всего случаев

101 518 229

Всего смертей

1 095 149

Текущие госпитализации

35 881

АКТУАЛЬНОСТЬ

Анализ влияния вакцинации

- Huang C. et al. Correlation between vaccine coverage and the COVID-19 pandemic throughout the world: Based on real-world data. – 2022.
- Chen X. et al. Impact of vaccination on the COVID-19 pandemic in US states. – 2022.

Прогнозирование заболеваемости

- Kumar Y. et al. Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Ten Nations' COVID-19 Pandemic. – 2022.
- Moftakhar L., Mozhgan S., Safe M. S. Exponentially increasing trend of infected patients with COVID-19 in Iran: a comparison of neural network and ARIMA forecasting models. – 2020.

Вероятностное прогнозирование

- Li X. et al. Probabilistic solar irradiance forecasting based on XGBoost.– 2022.

ЦЕЛИ И ЗАДАЧИ

Создание инструмента вероятностного прогнозирования развития пандемии Covid-19 на основе исторических данных по заболеваемости и статистики по вакцинации

- Анализ влияния вакцинации на заболеваемость;
- Прогнозирование заболеваемости на основе исторических данных;
- Вероятностное прогнозирование заболеваемости в зависимости от количества вакцинированных людей с использованием различных алгоритмов машинного обучения и статистических методов;
- Оценка точности работы моделей и сравнение моделей

ЭТАПЫ ИССЛЕДОВАНИЯ

Анализ литературы:
- поиск подходящих данных;
- обзор существующей литературы;
- обзор существующих моделей прогнозирования временных рядов

ШАГ 1

Исследование влияния вакцинации:
- на заболеваемость;
- на смертность

ШАГ 2

Подготовка данных:
- ресэмплирование по неделям;
- обработка пропущенных значений;
- проверка на стационарность;
- разделение на обучающую и тестовые выборки

ШАГ 3

Подготовка моделей для обучения:
- самостоятельное построение модели экспоненциального сглаживания;
- поиск готовых моделей;
- подбор гиперпараметров модели;
- адаптация моделей **машинного обучения** для прогнозирования временных рядов

ШАГ 4

Оценка точности работы модели:
- подсчет метрик MAPE, MAE, RMSE

ШАГ 5

Прогнозирование:
- прогнозирование заболеваемости с использованием данных по вакцинации и без;
- вероятностное прогнозирование

ШАГ 6

ДАННЫЕ

Заболеваемость

Вакцинация

date	Аббревиатура штата	Количество заболевших	Количество людей, вакцинированных хотя бы одной дозой	Количество людей, полностью вакцинированных
2020-03-09	AL	0.0	0.0	0.0
2020-03-10	AL	0.0	0.0	0.0
2020-03-11	AL	3.0	0.0	0.0
2020-03-12	AL	1.0	0.0	0.0
2020-03-13	AL	4.0	0.0	0.0
...	...	...	...	...
2023-03-01	AL	3714.0	0.0	0.0
2023-03-02	AL	0.0	639.0	559.0
2023-03-03	AL	0.0	0.0	0.0
2023-03-04	AL	0.0	0.0	0.0
2023-03-05	AL	0.0	0.0	0.0

АНАЛИЗ ВЛИЯНИЯ ВАКЦИНАЦИИ

Взаимная корреляция — это способ измерения степени сходства между временным рядом и запаздывающей версией другого временного ряда.

Этот тип корреляции полезен для расчета, потому что он может показать, предсказывают ли значения одного временного ряда будущие значения другого временного ряда. Другими словами, он может сказать нам, является ли один временной ряд опережающим индикатором для другого временного ряда.

Взаимная корреляция для каждой задержки между двумя временными рядами рассчитывается через кросс-корреляционную функцию:

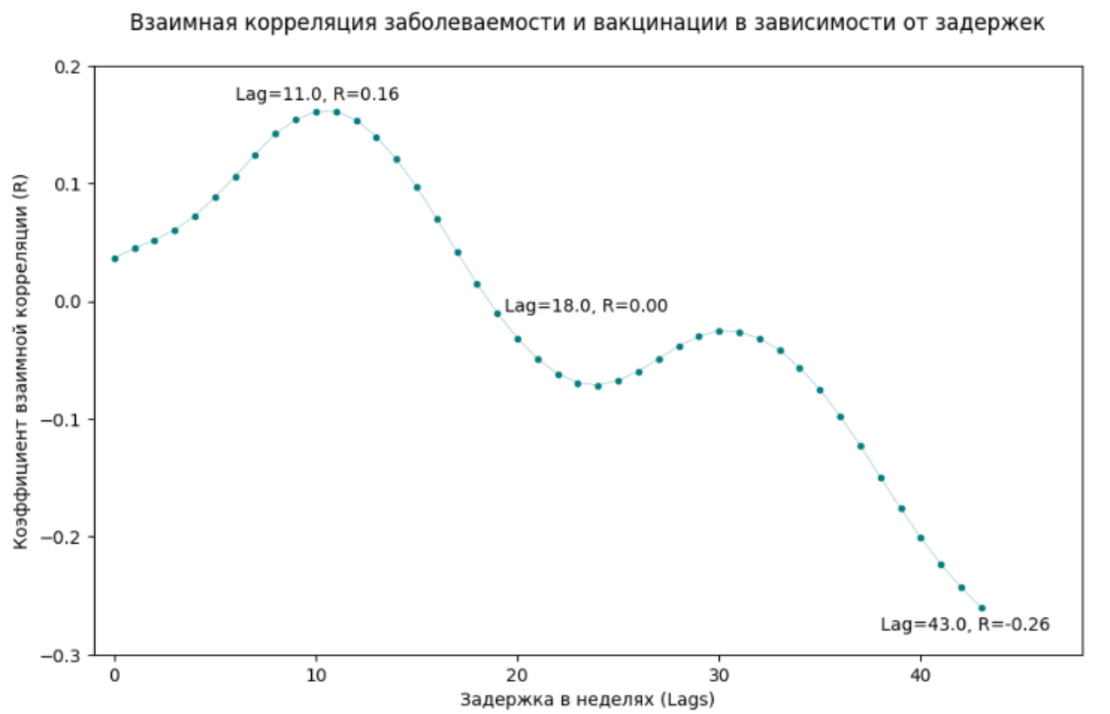
$$\rho_{xy}(l) = \frac{\sum_{i=0}^{N-1}(x_i-\bar{x})*(y_{i-l}-\bar{y})}{\sqrt{\sum_{i=0}^{N-1}(x_i-\bar{x})^2}\sqrt{\sum_{i=0}^{N-1}(y_{i-l}-\bar{y})^2}} = \frac{cov(x,y)}{\sqrt{\sigma_x^2\,\sigma_y^2}}$$

$cov(x,y)$ – ковариация

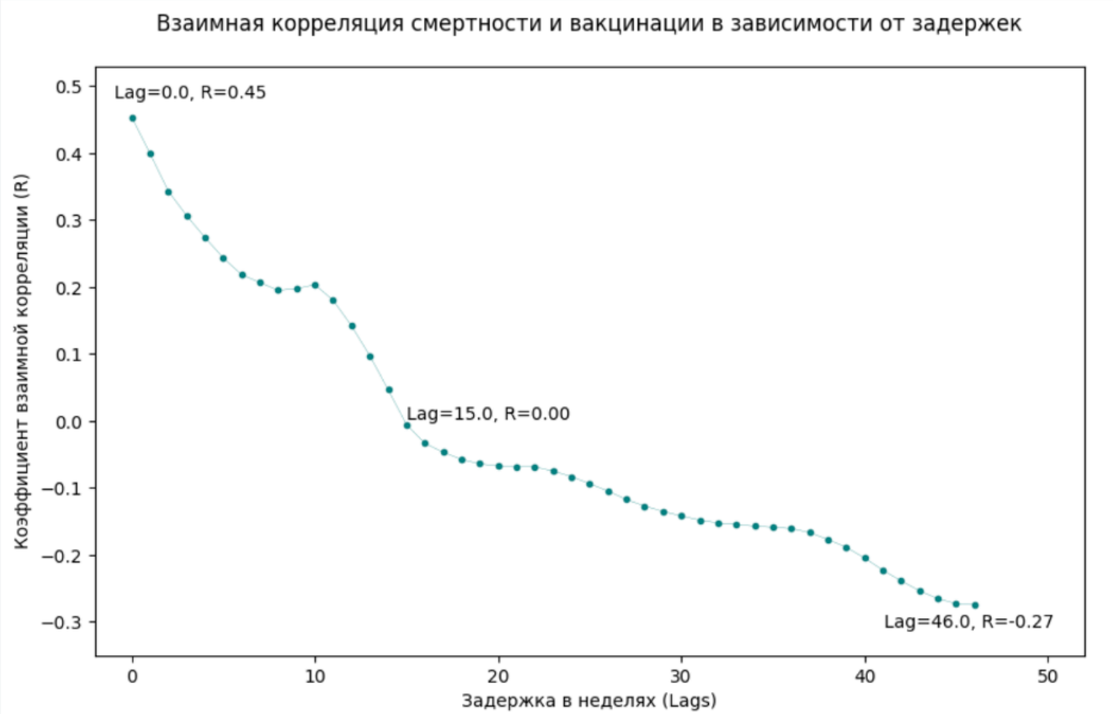
σ_x^2 и σ_y^2 – дисперсия

АНАЛИЗ ВЛИЯНИЯ ВАКЦИНАЦИИ

Коэффициент корреляции, равный **-1**, показывает идеальную **отрицательную корреляцию**.
Коэффициент корреляции, равный **1**, показывает идеальную **положительную корреляцию**.



Задержка (lag), в неделях	Минимум	Максимум
11	-	0.16
43	-0.26	-



Задержка (lag), в неделях	Минимум	Максимум
0	-	0.45
46	-0.27	-

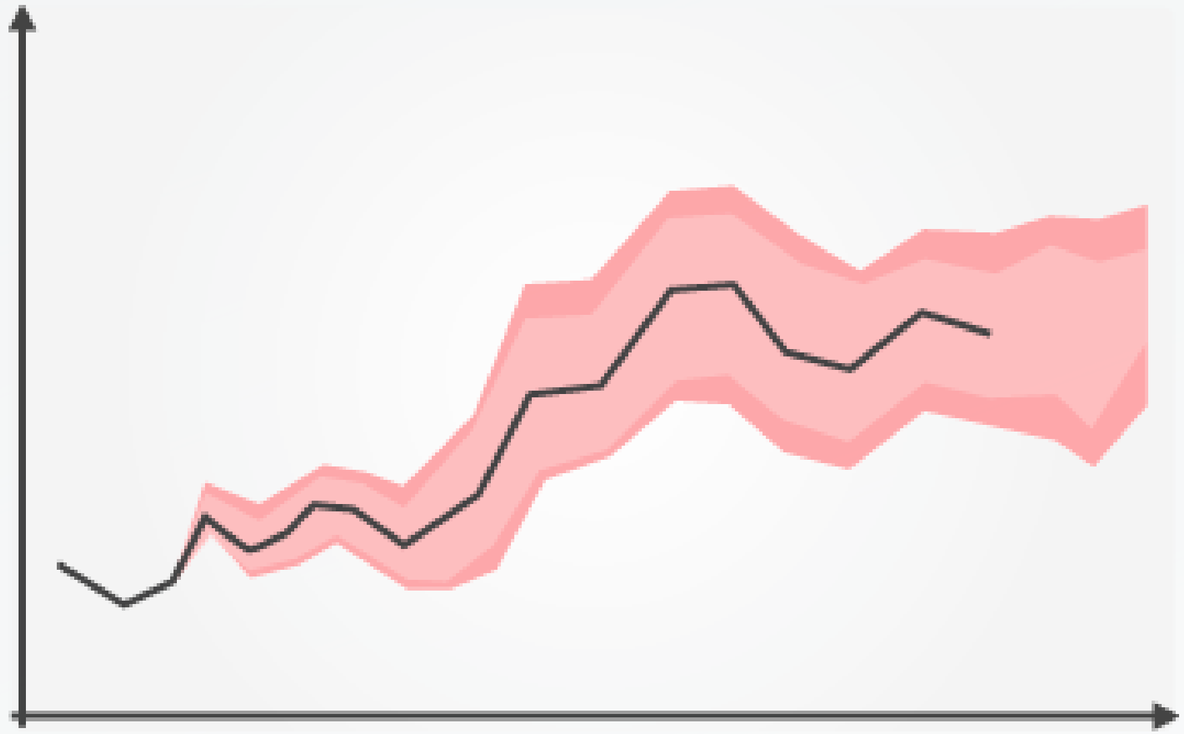
ВИДЫ ПРОГНОЗИРОВАНИЯ

Точечное прогнозирование

Точечный прогноз дает только детерминированный результат, он содержит ограниченную информацию о случайном и флуктуирующем процессе.

Вероятностное прогнозирование

Вероятностное прогнозирование позволяет прогнозировать ожидаемое распределение результата, а не одно будущее значение. Этот тип прогнозирования предоставляет гораздо более богатую информацию, поскольку он сообщает диапазон вероятных значений, в который может попасть истинное значение. Для этого используется прогнозирование на основе бутстрэпных выборок остатков (ошибок) прогноза.



МЕТОДЫ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

СТАТИСТИЧЕСКИЕ МЕТОДЫ

- Авторегрессионное интегрированное скользящее среднее (ARIMA)
- Методы экспоненциального сглаживания

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

- k-ближайших соседей
- Градиентный бустинг
- Случайный лес

ARIMA

Модели ARIMA – это класс моделей, способных прогнозировать как стационарные, так и нестационарные временные ряды на основе исторических данных.

Модель ARIMA характеризуется 3 переменными и обозначается как ARIMA(p,q,d):

- **p – порядок авторегрессии (AR)**, который позволяет добавить предыдущие значения временного ряда;
- **d – порядок интегрирования (I)**, который указывает порядок разностей, необходимых для формирования стационарного временного ряда;
- **q – порядок скользящего среднего (MA),** который позволяет установить ошибку модели как линейную комбинацию наблюдавшихся ранее значений ошибок.

Вместе эти три параметра позволяют учитывать сезонность, тренд и шум.

Модель ARIMA описывается уравнением:

$$y'_t = I + \beta_1 y'_{t-1} + \beta_2 y'_{t-2} + .. + \beta_n y'_{t-n} + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + .. + \varphi_n \varepsilon_{t-n}$$

$y'_t = y_t - t_{t-d}$ – разностный ряд порядка d

I - разностные значения ряда порядка d

β – коэффициент задержки (lag)

ε_t – ошибка прогноза

Во время обучения ARIMA производит оценку коэффициентов β и φ для заданных p, d, q.

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

Экспоненциальное сглаживание − это средневзвешенное значение прошлых данных, при этом последним точкам данных придается больший вес, чем более ранним точкам данных. Веса экспоненциально затухают по направлению к более ранним точкам данных.

Простое экспоненциальное сглаживание:

$$F_{n+1} = \alpha\, y_n + (1-\alpha) F_n \qquad F_n = \alpha\,(\,y_{n-1} + \alpha\,(1-\alpha)\,y_{n-2} + ..\,)$$

Двойное экспоненциальное сглаживание Холта:

$$F_{n+1} = L_n + T_n$$

$$L_n = \alpha\, y_n + (1-\alpha)(L_{n-1} + T_{n-1})$$ - среднее значение или «уровень» временного ряда

$$F_{n+1} = L_n + h\, T_n$$

$$T_n = \beta\,(L_n + L_{n-1}) + (1-\beta)\, T_{n-1}$$ - компонент тренда временного ряда

Тройное экспоненциальное сглаживание Холта-Винтерса:

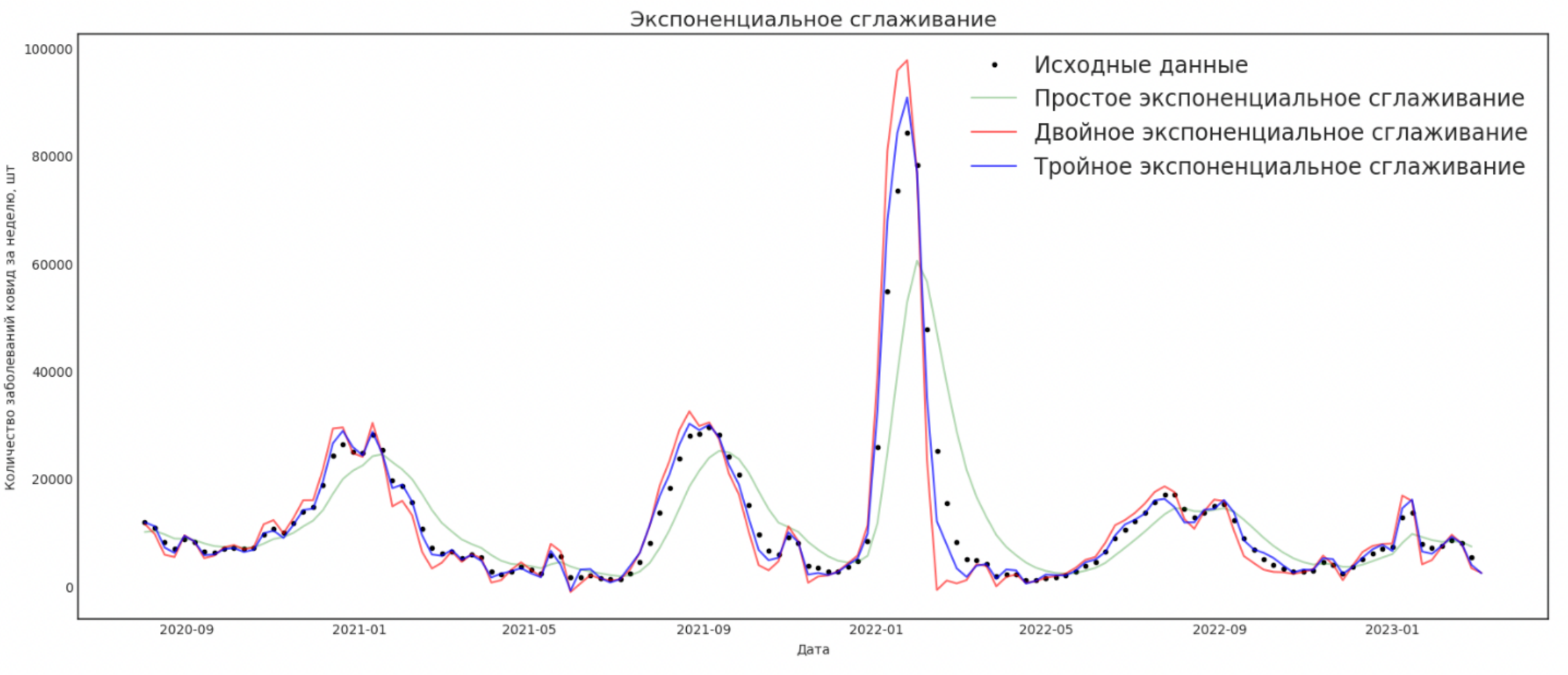
$$F_{t+h} = (L_t + h\, T_n) S_{t+h-p}$$

$$L_t = \frac{\alpha\, y_t}{S_{t-p}} + (1-\alpha)(L_{t-1} + T_{t-1})$$ - среднее значение или «уровень» временного ряда

$$T_t = \beta\,(L_t - L_{t-1}) + (1-\beta)\, T_{t-1}$$ - компонент тренда временного ряда

$$S_t = \gamma \left(\frac{y_t}{L_t}\right) + (1-\gamma)\, S_{t-p}$$ - сезонный компонент

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ



К-БЛИЖАЙШИХ СОСЕДЕЙ

Метод k-ближайших соседей (k Nearest Neighbors, или kNN) – алгоритм для решения задач регрессии. На интуитивном уровне суть метода проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься.

Для каждой новой точки данных, KNN находит k ее наиболее похожих примеров, называемых ближайшими соседями, в соответствии с метрикой расстояния (расстояние Чебышёва, Манхэттенское расстояние, Евклидово расстояние).

Формула Евклида:

$$\sqrt{\sum_{x=1}^{n}(f_x^i - q_x)^2}$$

f_x^i - точка из обучающей выборки

q_x - пример точки

Объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны:

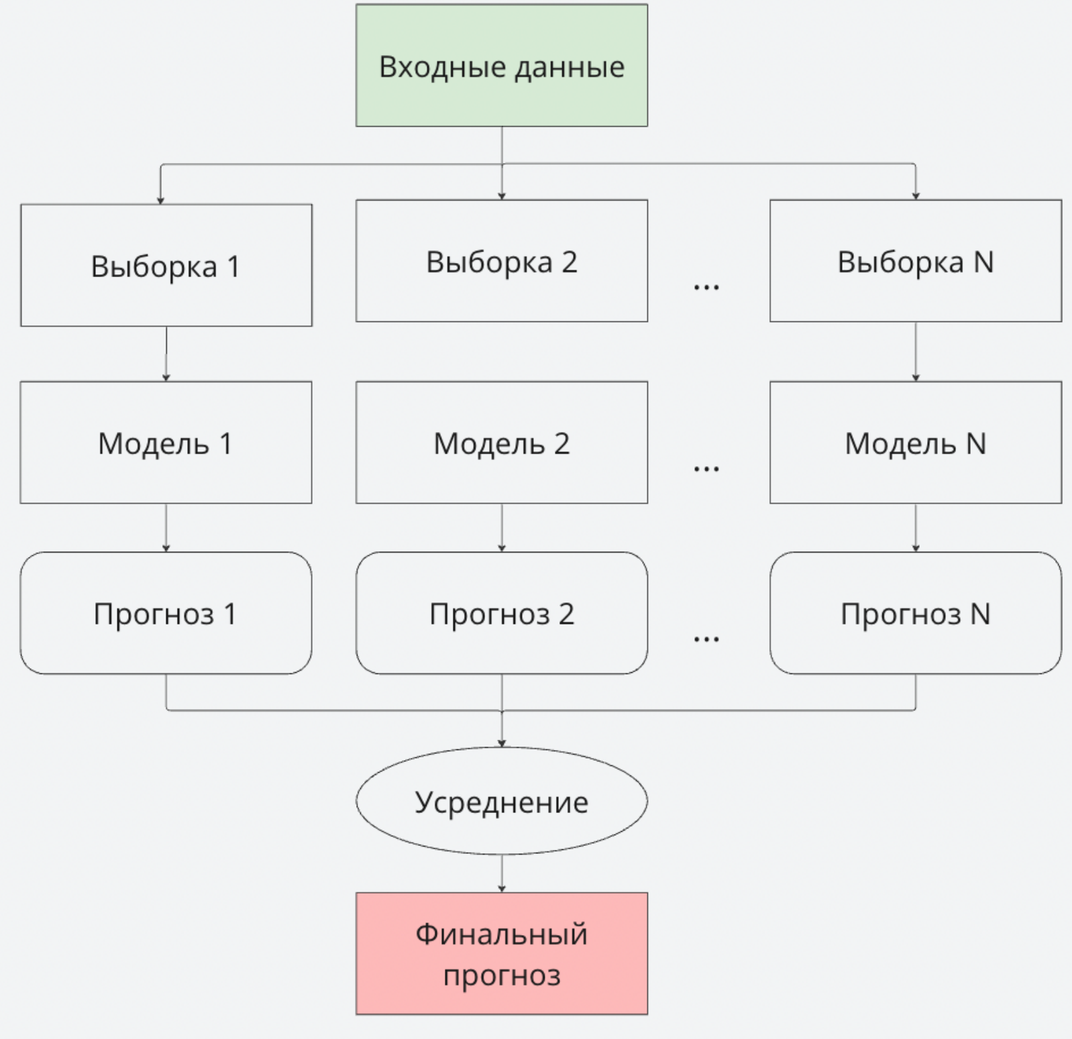
$$\hat{y} = mean\,(w_1 * y_1, w_2 * y_2, \dots w_k * y_k)$$

$$\omega_i = \frac{e^{-d(x,n_i)}}{\sum_{i=1}^{k} e^{-d(x,n_i)}}$$ - вес

СЛУЧАЙНЫЙ ЛЕС

Алгоритм случайного леса (Random Forest) – алгоритм машинного обучения, который строится на основе ансамбля решающих деревьев.

Дерево решений – это алгоритм машинного обучения, который стоится на основе иерархического объединения логических правил вида «если .., то ..».

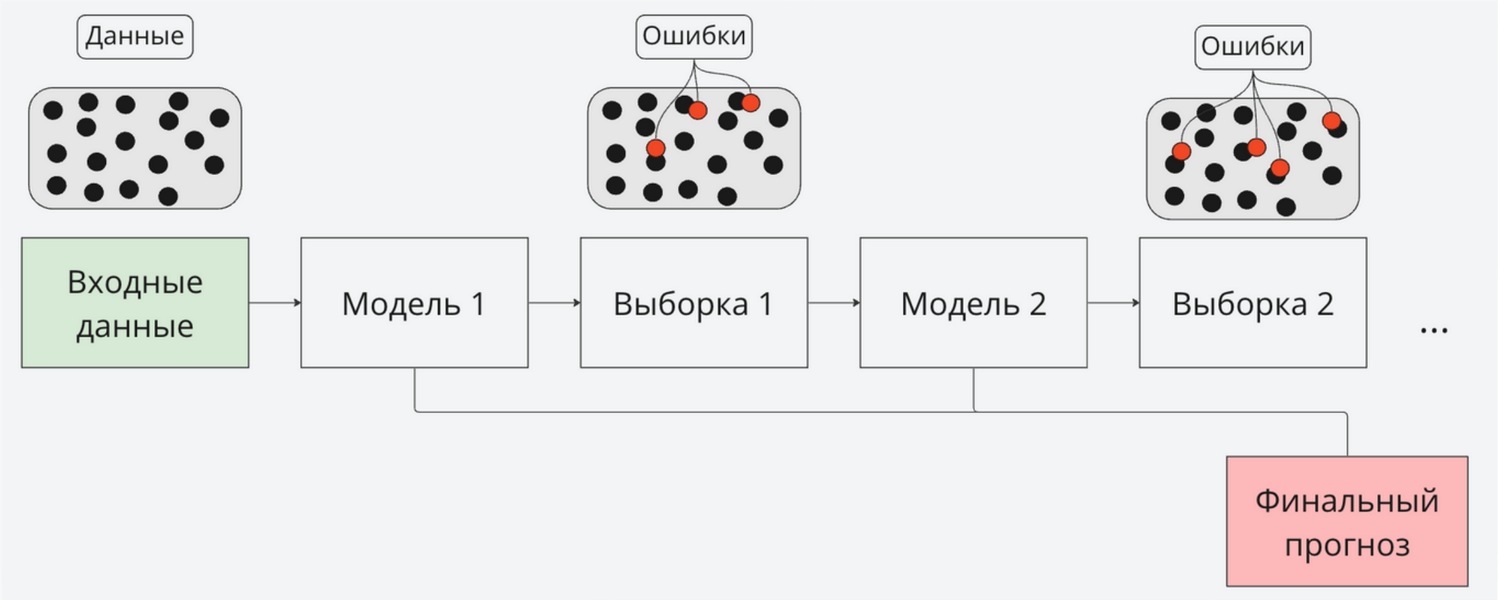


ГРАДИЕНТНЫЙ БУСТИНГ

Градиентный бустинг — это метод машинного обучения, представляющий собой линейную аддитивную модель, состоящую из ансамбля слабых моделей прогнозирования.

Градиентный бустинг включает в себя выявление недостатков слабых моделей и последовательное построение окончательной модели ансамбля с использованием функции потерь, оптимизированной с помощью градиентного спуска.

Основная идея градиентного бустинга заключается в постоянном добавлении слабых деревьев с разным весом в набор. Деревья в наборе должны максимально приближаться к остаткам предыдущего прогноза.



ОЦЕНКА ТОЧНОСТИ МОДЕЛЕЙ

MAPE – средняя абсолютная ошибка в процентах

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} * 100\%$$

MAE – средняя абсолютная ошибка

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MedAPE – медианная абсолютная ошибка в процентах

$$MedAPE = median\,(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|)$$

RMSE – среднеквадратичная ошибка

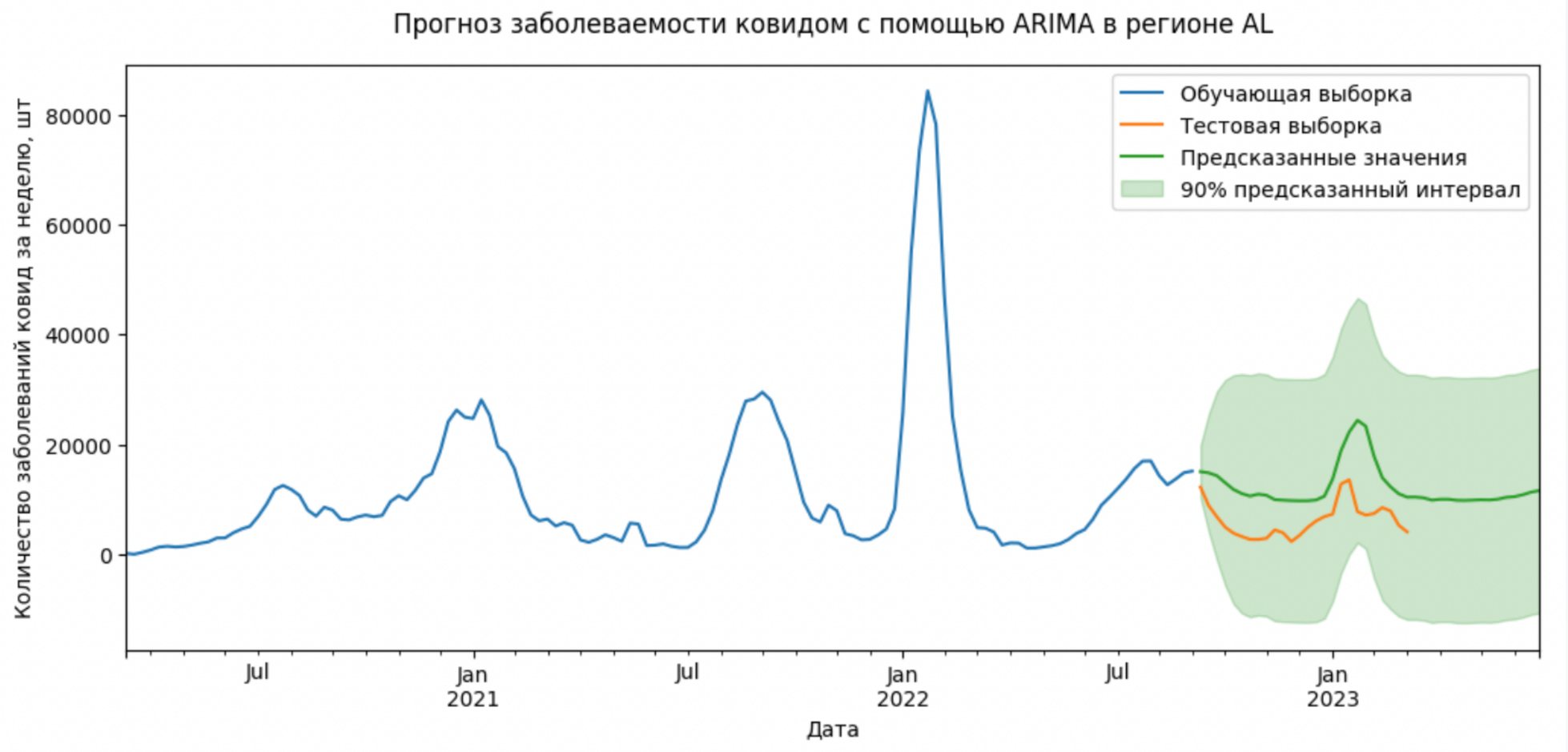
$$RMSE = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}$$

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

МЕТРИКИ

- MAPE = 85%
- MedAPE = 66%
- MAE = 4213
- RMSE = 5027

ПРОГНОЗ С ПОМОЩЬЮ ARIMA



МЕТРИКИ

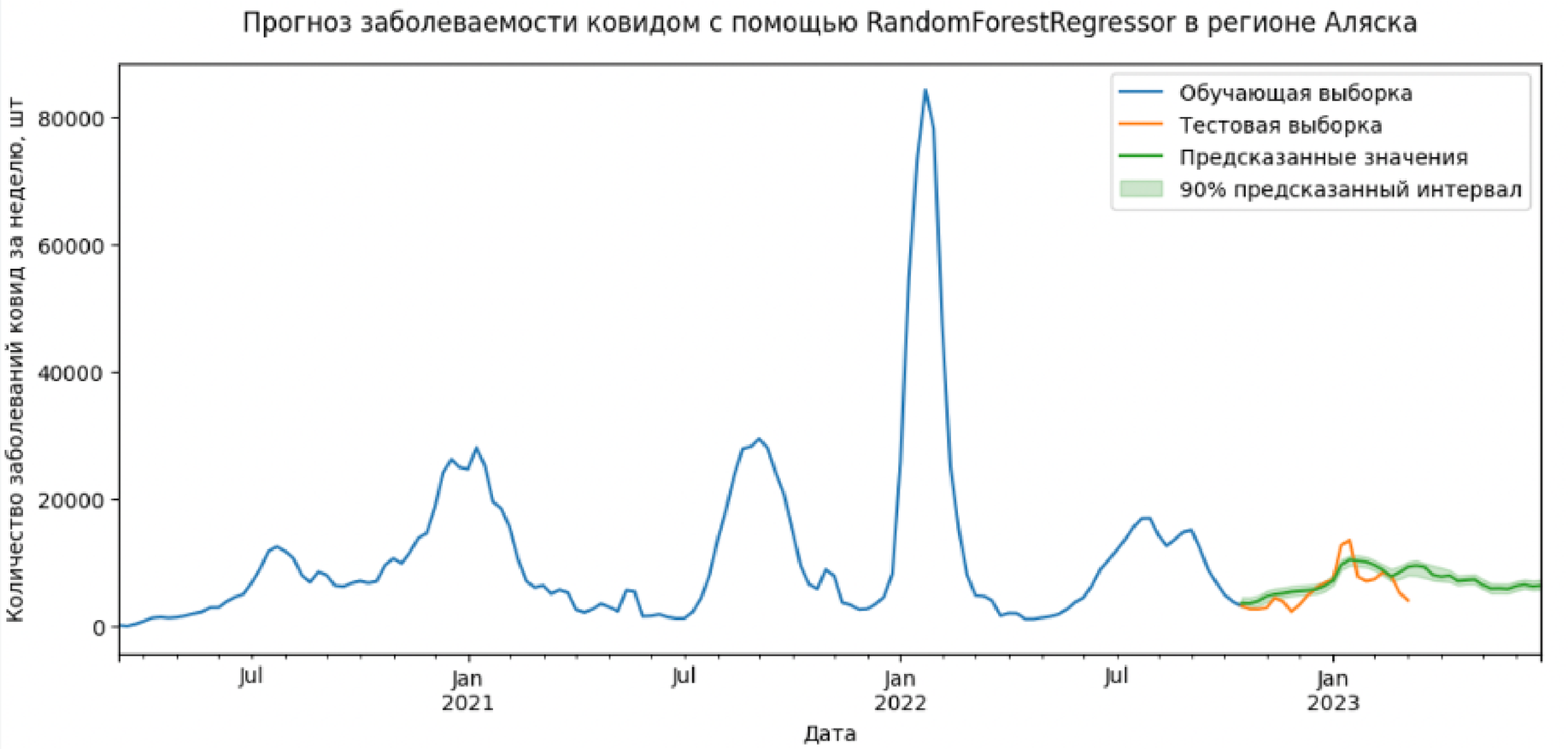
- MAPE = 132%
- MedAPE = 125%
- MAE = 6387
- RMSE = 7366

К-БЛИЖАЙШИХ СОСЕДЕЙ

МЕТРИКИ

- MAPE = 55%
- MedAPE = 43%
- MAE = 2711
- RMSE = 2884

СЛУЧАЙНЫЙ ЛЕС



МЕТРИКИ

- MAPE = 36%
- MedAPE = 29%
- MAE = 1711
- RMSE = 2186

ГРАДИЕНТНЫЙ БУСТИНГ

МЕТРИКИ

- MAPE = 31%
- MedAPE = 22%
- MAE = 1854
- RMSE = 2983

СРАВНЕНИЕ МОДЕЛЕЙ

Модель	Экспоненциальное сглаживание	ARIMA	k-ближайших соседей	Случайный лес	XGBoost
Ошибка MAPE	102%	154%	90%	87%	58%
Ошибка MAPE с использованием статистики по вакцинации	-	-	79%	80%	53%

ЗАКЛЮЧЕНИЕ

В данной работе оценивается возможность применения вероятностного моделирования для прогнозирования развития вирусных заболеваний. При этом использование дополнительной информации по вакцинации позволяет повысить точность прогноза, не смотря на слабое влияние вакцинации на заболеваемость.

Ошибка моделей варьируется от 30% до 130%, но предсказанные значения достаточно неплохо описывают характер временного ряда, его тренд и сезонность.

В результате работы был создан готовый инструмент на языке Python, который производит предварительную обработку данных и делает прогноз в будущее. При этом достаточно 50 недель, чтобы с точностью 70% предсказывать данные вперед на 4–5 недель.

Полученные в результате прогнозные модели развития пандемий могут оказаться полезным для принятия защитных мер и разработки плана действий во время таких пандемий, как Covid-19.

Выполненные задачи:

- ✓ Анализ корреляции вакцинации и заболеваемости
- ✓ Прогнозирование заболеваемости на основе исторических данных
- ✓ Вероятностное моделирование развития пандемии в зависимости от количества вакцинированных людей с использованием алгоритмов машинного обучения и статистических методов
- ✓ Оценка точности работы моделей и сравнение моделей

СПАСИБО ЗА ВНИМАНИЕ!