

ПРЕЗЕНТАЦИЯ

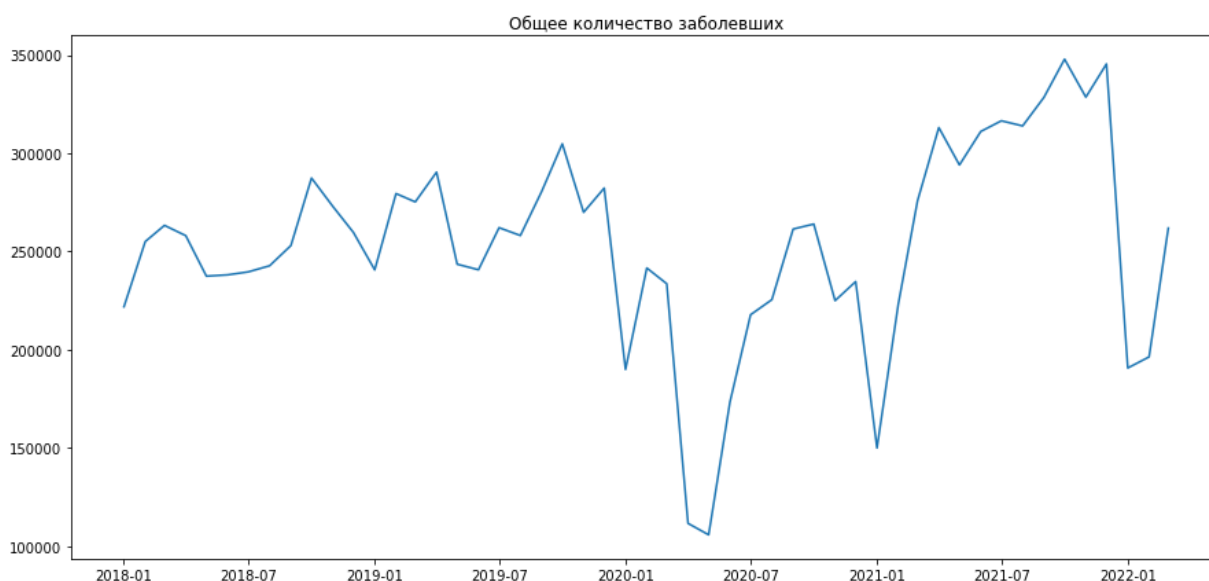
«Создание модели прогнозирования заболеваемости населения»

Для решения задачи был использован комбинированный подход:

1. На первом шаге произведена оценка цикличности общего количества заболевших,
2. так как стандартный подход (дифференцирование, стандартное отклонение и др.) с дальними периодами ухудшал метрики, был выбран подход краткосрочного прогнозирования из информационных данных за предыдущие три месяца

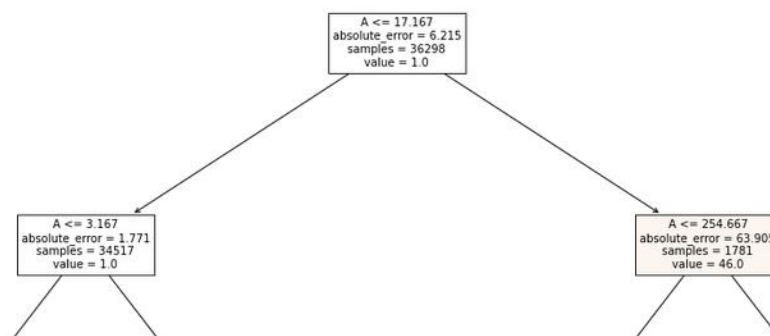
3. Тестирование производилось на аналогичном ближайшем периоде 2021 года с валидацией на апреле 2021 года. Окончательное решение о принятии модели происходило по текущему рейтингу на лидерборде. Модель может быть улучшена совместно с профессиональными эпидемиологами.

(схожесть визуальной динамики по графику заболеваемости).



4. Из-за несбалансированности в данных, сделана группировка по среднему за предыдущие месяцы.

Поиск крайних значений бинов произведен с помощью DecisionTreeRegressor



Т.е. взяты крайние бины до 3,167, между 3,167 и 254,667, более 254,667.

Модель “самонастраивается” – ищет разделение автоматически

5. Каждый бин обучался по "своей" модели.

А также дерево использовалось для оценки значимости признаков, например:

```
# Бин 1
tree_bin_1 = DecisionTreeRegressor()
tree_bin_1.fit(X_bin_1, y_bin_1)
pd.Series(np.round(tree_bin_1.feature_importances_ * 100, 3), index=X_bin_1.columns)
```

PREV_PATIENT_ID_COUNT	5.246
AVG_PATIENT_ID_COUNT	9.414
PATIENT_MKB_ADRES_COUNT	24.185
PATIENT_MKB_SEX_COUNT	11.260
PATIENT_MKB_AGE_COUNT	15.549
PATIENT_MKB_COUNT	8.702
PATIENT_ADRES_COUNT	5.244
PATIENT_SEX_COUNT	1.709
PATIENT_AGE_COUNT	4.569
COUNT_MKB_VS_MKB_ADRES	14.124

dtype: float64

Для запуска модели необходимо ввести константы:

RANDOM_STATE = 0 # фиксирование случайного числа

PREDICT_MONTH = 4 # номер месяц для прогноза

PREDICT_YEAR = 2022 # год прогноза

TRAIN_PERIOD = 3 # кол-во месяцев, предшествующих прогнозу, на которых будет обучаться модель (подобрано опытным путем)

Технические особенности:

Python, LinearRegression, KNN, DecisionTreeRegressor, MaxAbsScaler, PolynomialFeatures, pipeline

Возможно, снижение точности алгоритма детектирует эпидемиологические аномалии.