Финальный проект по теме: “Предсказание временных рядов”

Выполнил: Запсельский Виктор
Проверили: Иван Еремеев
Дарья Чиркина
Евгений Бокарев

образование

1) Condition, relevance of the task and my solution

Актуальность выбранной темы:

Думаю для всего мира на 2021 год проблема распространения COVID-19 является одной из главных проблем. Из-за большой заболеваемости в мире в 2019 году страны были вынуждены вводить локдауны по всему миру, экономика мира терпела убытки и, безусловно, хотелось бы “сгладить” убытки. Поэтому, чтобы не допустить неожиданных результатов развития заболеваемости было бы хорошо знать динамику распространения заболевания на ближайшее будущее.

Лично для меня предсказание временных рядов является очень интересной задачей, так как я воспринимаю это как некоторое “соревнование”: в ходе решения задачи нужно уметь выбирать правильную модель для конкретного случая, правильно её оптимизировать и улучшать качество модели по выбранным метрикам с каждым baseline’ом. А участвовать в соревнованиях и побеждать я люблю с детства :)

Практические вопросы, которые были исследованы в ходе работы:

1)Был проведён первичный анализ данных

а) построены графики заражений/ смертности для ТОП 5 стран, которые имеют самый высокий показатель заражения, и сделаны выводы по этим графикам
б) сделана визуализация данных на карте с помощью библиотеки keplergl

2) Реализованы две модели предсказания временных рядов, а также выполнен подбор гиперпараметров для каждой из модели:

1-ая с помощью библиотеки statsmodels (модель Arima)
2-ая с помощью библиотеки facebook.Prophet (модель Prophet)

3) Выполнен прогноз кол-ва заражений/смертей в день по миру и по России на 3, 6,9, 12 и 15 дней соответственно.

4) Сделана визуализация данных для полученного прогноза

5) Проведено сравнение двух моделей для конкретных случаев

Task and data description:

Task: Необходимо сравнить качество прогноза по изученным методам прогнозирования временных рядов по метрикам MAPE/MAE. В качестве сравнения использовать столбцы ConfirmedCases, Fatalities из предложенного датасета. Проверить прогноз отдельно для России, отдельно для всего мира.

Данные для прогнозирования: COVID 19

Описание данных:

- датасет train содержит информацио о распространении короновирусной инфекции за 4 месяца от 2020-01-22 до 2020-05-15
- Country_Region - название страны
- Province_State - штат/провинция. У одной страны может быть несколько штатов.
- ConfirmedCases - общее кол-во подтвержденных случаев по штату/провинции.
- Fatalities - общее кол-во смертей по штату/провинции. Если в Province_State - значение Nan, то Fatalities рассчитывается по Country_Region

Дополнительно были созданы следующие таблицы и добавлены в них новые features. В дальнейшем все операции производились с этими тремя таблицами:

1) total_by_country - таблица с суммарными значениями кол-ва заболевших/смертей по каждой стране
total_ConfirmedCases - суммарное кол-во заболевших в стране
total_Fatalities - суммарное кол-во смертей в стране

2) df_by_state - таблица с данными по каждому штату
Fatalities_per_day - кол-во смертей по штату в день
ConfirmedCases_per_day - кол-во заражений по штату в день

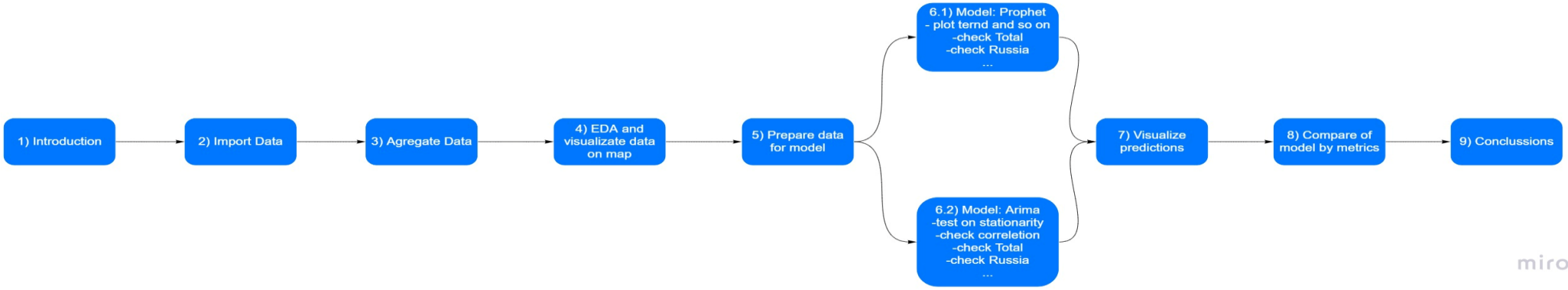
3) df_by_country - таблица с данными агрегированными по каждой стране
FPD_by_countr - кол-во смертей по стране в день
CPD_by_countr - кол-во заражений по стране в день
F_by_countr - кол-во смертей по стране (учитывая кол-во предыдущих дней)
C_by_countr - кол-во заражений по стране (учитывая кол-во предыдущих дней)

Solution:

В качестве решения данной задачи мной были выбраны 2 модели прогнозирования : ARIMA and Prophet.
Прогноз выполнялся для следующих временных рядов: количества зараженных/смертей за всё время по миру, количества зараженных/смертей за день по миру и аналогично для России.
Метрики сравнения: MAE and MAPE.

Ход выполнения проекта:

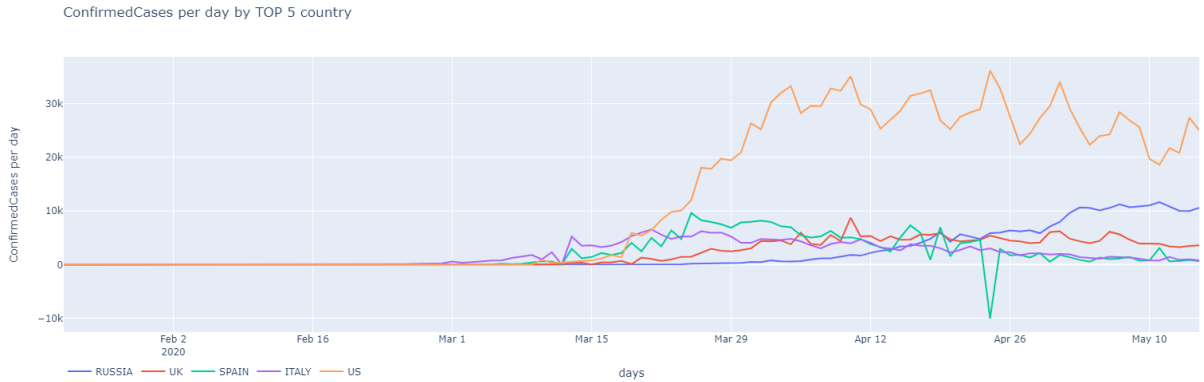
1) предобработка данных с помощью библиотеки Sqlite: созданы новые таблицы, добавлены в них новые features
2) первичный анализ данных: визуализация графиков, выявление зависимостей, визуализация данных на maps.
3) построение модели ARIMA:
    а) разложение рядов на компоненты
    б) проверка рядов на стационарность
    в) графики автокорреляции и частичной корреляции
    г) выполнение predict'а и расчёт ошибок MAE и MAPE
4) Построение модели PROPHET:
    а) подбор гиперпараметров модели, используя валидацию
    б) выполнение predict'а и расчёт ошибок MAE и MAPE
5) Визуализация предсказанных данных на графиках и визуализация ошибок MAE/MAPE двух моделей.
7) Выводы о качестве предсказаний моделей

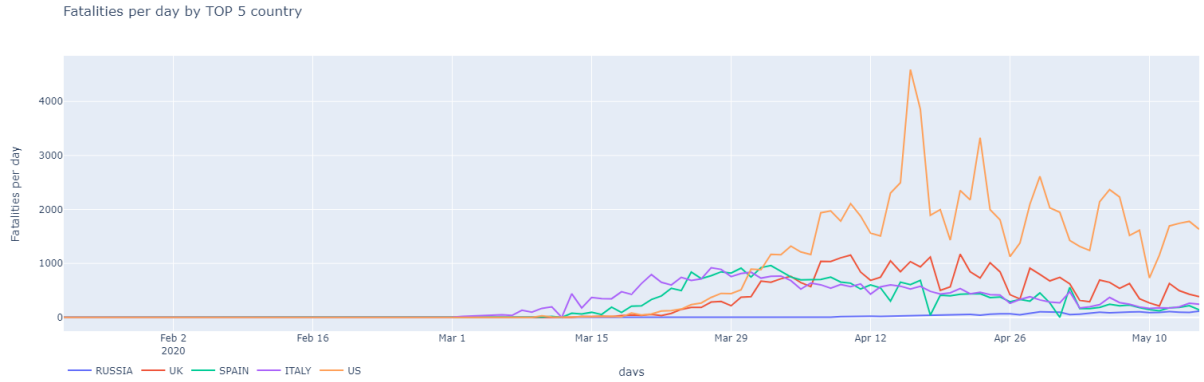


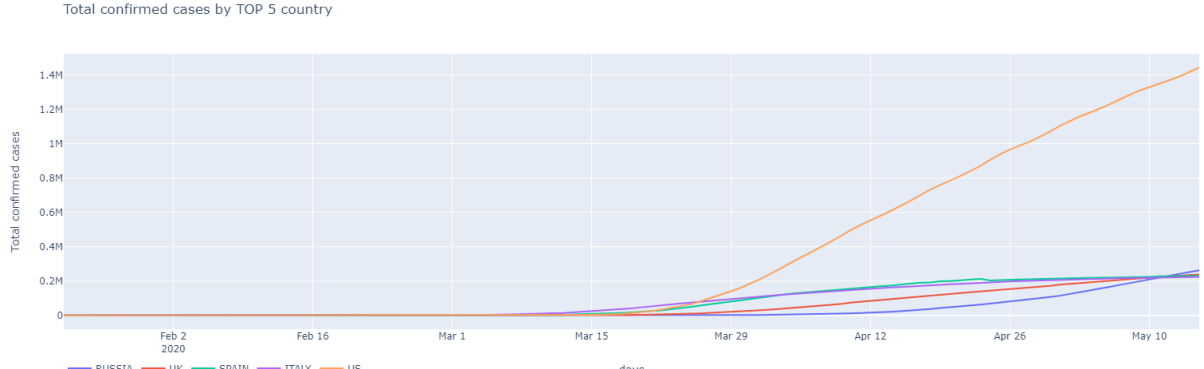
Ссылка на Google Collab: SOLUTION

2) Exploratory Data Analysis

Charts:







Conclusions:

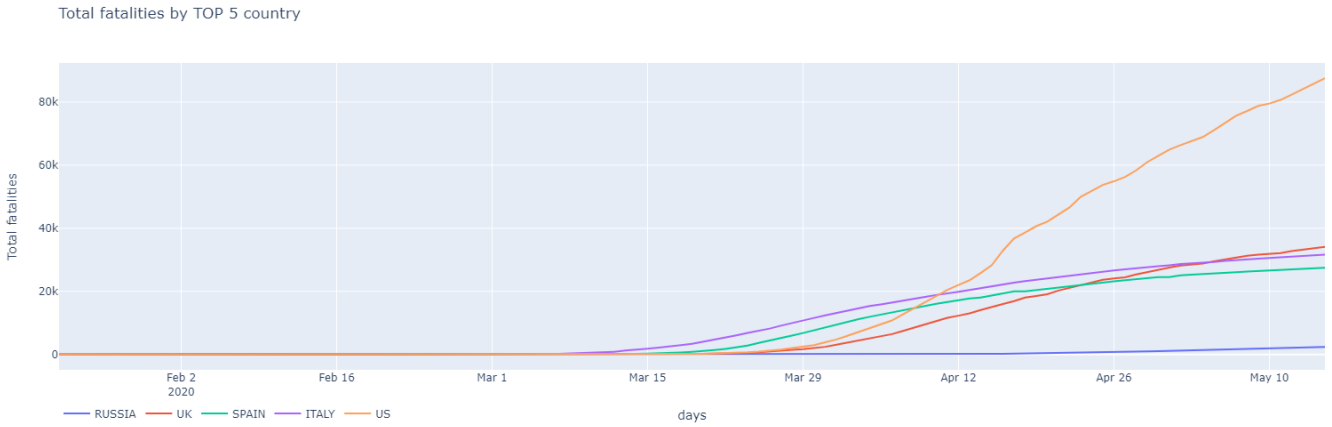
ТОП 5 стран с максимальным показателем confirmed cases:
1) US 2) Russia 3) UK 4) Italy 5) Spain

- США вышли на первое место по показателю количеству заражений в день с 21 марта
- Россия вышла на второе место по показателю количеству заражений в день с 19 апреля
- Число количеству заражений в день в Англии, Испании и Италии находится в одном диапазоне и ниже, чем в России
-графики confirmed cases per day имеют нелинейную зависимость

-максимальный показатель смертей в день был зафиксирован в США 16 апреля и составил 4591
- Число летальных исходов в Англии, Испании и Италии находится в одном диапазоне
- в России наблюдается самое низкое число смертей из выбранных стран
- графики fatalities per day имеют нелинейную зависимость

- Количество новых заболевших в США значительно увеличивается и заметно превосходит количество зараженных в других странах
- Россия ~ с 12 мая на втором месте по общему кол-ву заражений
-динамика общего числа заражений в данных странах имеет линейную зависимость

Charts:



Conclusions:

- Интересно, что количество летальных исходов в день и общее число летальных исходов в России очень низкое по сравнению с другими странами. Хотя, как было сказано выше, Россия - на втором месте по количеству заражений в день и на втором месте по общему числу заболевших.

Наверное, Россияне самый живучий народ :)

Так, в России:

- кол-во смертей за всё время: ~2400 человек
- кол-во заражений за всё время: ~260k человек

В Англии:

- кол-во смертей за всё время: ~35000 человек
- кол-во заражений за всё время: ~230k человек

- Слева представлена визуализация данных на карте: используя geocoder Nominatim получены координаты стран, далее с помощью библиотеки KeplerGl данные отображены на map.

Можно настроить различные фильтры и отследить динамику заражения на карте.

3) Making predictions.
Model: ARIMA

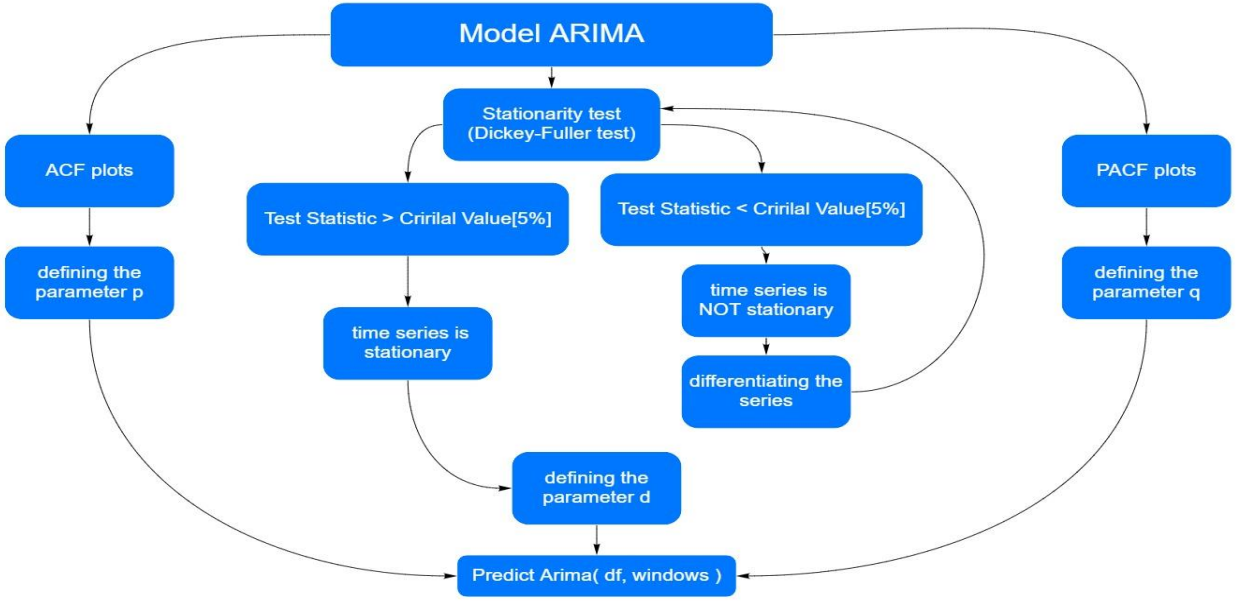
Model: ARIMA

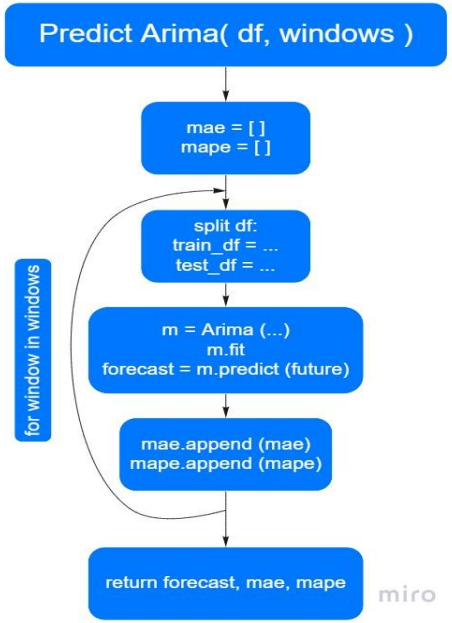
Основные компоненты модели (p,d,q):

AR (Autoregressive) - компонента авторегрессии (p).
Для нахождения данного параметра будем использовать Plot the autocorrelation function из библиотеки statmodels. В параметрах plot_acf передаё м кол-во лагов на графике, после чего смотрим на plot и считаем количество лагов, которые имеют высокую корреляцию. Кол-во таких лагов = p.

I (Integrated) - компанента интегрирования (d).
Выбор данного параметра будем производить используя тест Дики-Фуллера. Значение параметра зависит от того, сколько раз необходимо продеффиринцировать ряд для того, чтобы он принял стационарный вид. В случае, если ряд уже стационарный, то d = 0 и модель будет называться ARMA.

MA (Moving Average) - компонента скользящего среднего (q).
Для нахождения данного параметра будем использовать Plot the partial autocorrelation function из библиотеки statmodels (график частичной автокорреляции). Подбор данного параметра аналогичен параметру p. В параметрах plot_pacf передаём кол-
во лагов на графике, после чего смотрим на plot и считаем количество лагов, которые имеют высокую корреляцию. Кол-во таких лагов = q.

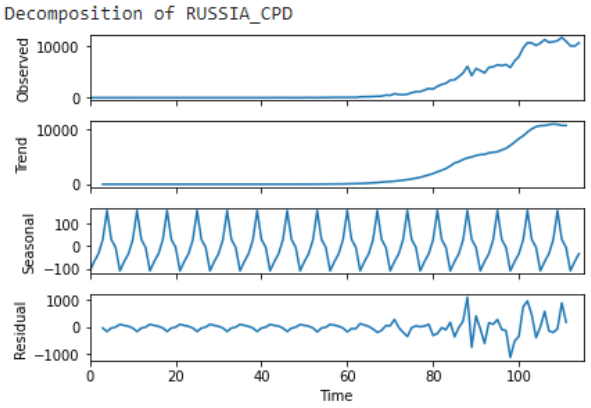


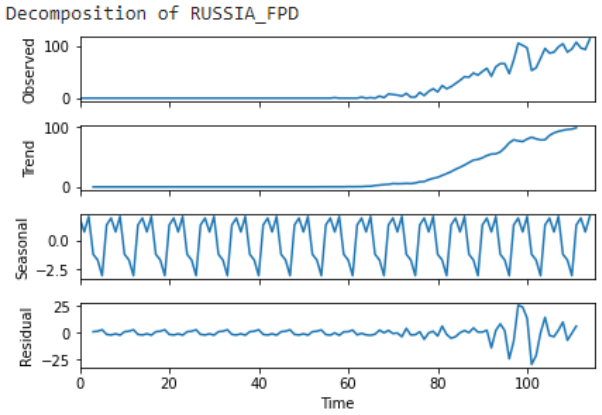


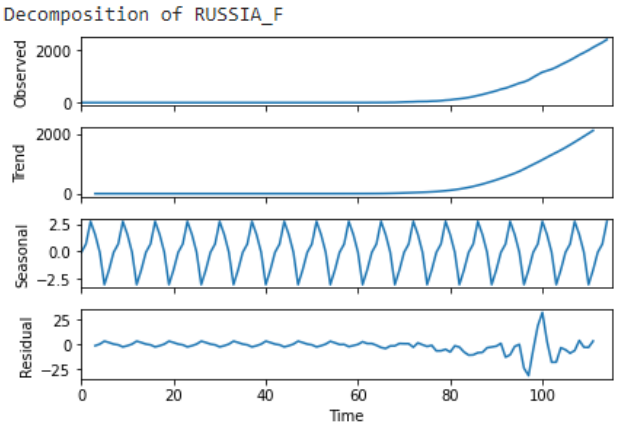
Model: ARIMA

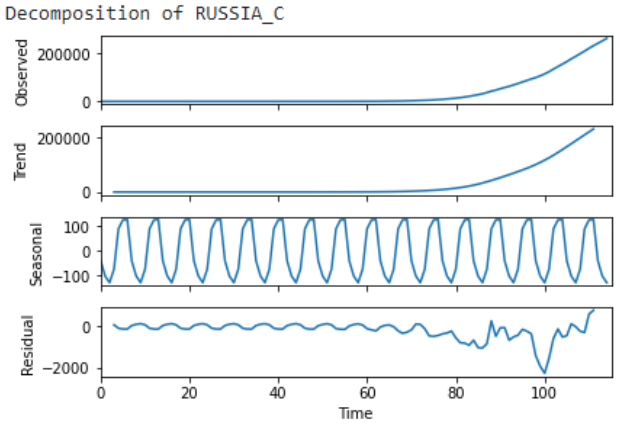
Decomposition: выполним разложение рядов TOTAL and RUSSIA на три компоненты: тренд, сезонность и остатки.

For Russia:

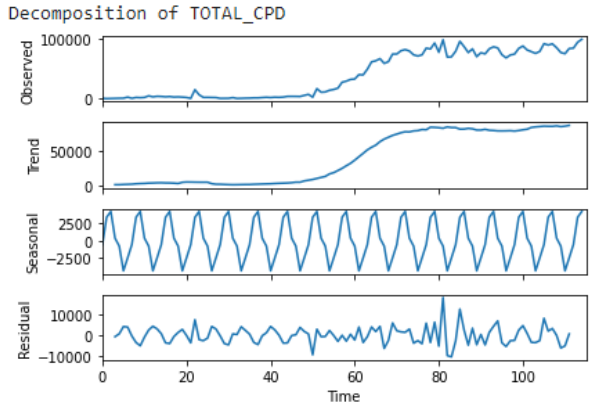


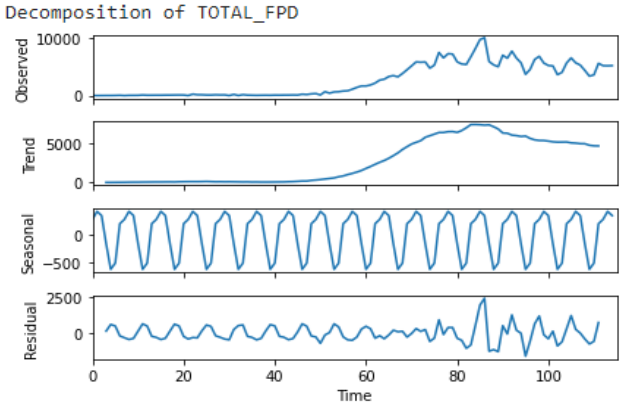


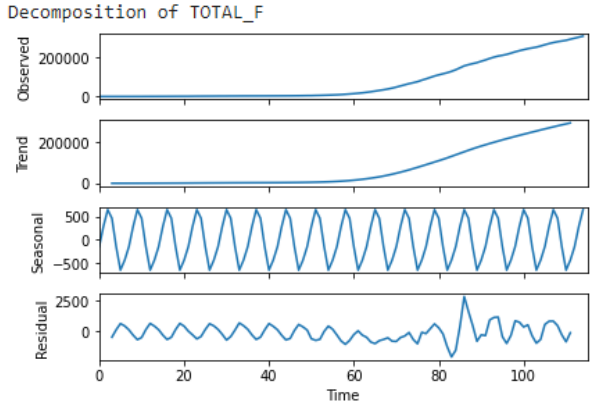


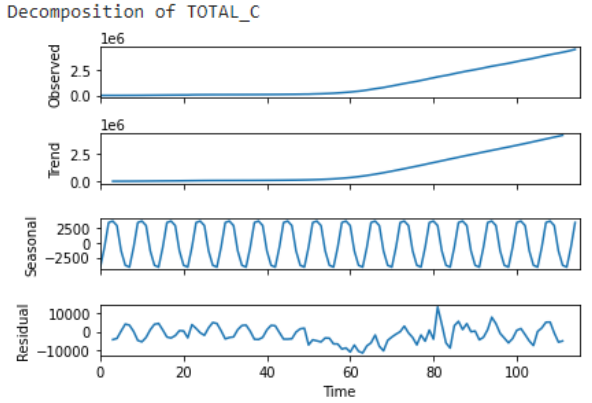


For Total:









Conclusion: из анализа декомпозиции рядов можно заметить, что практически все ряды имеют восходящий тренд. Исключение является ряд TOTAL_FPD, где тренд сменил своё направление и теперь является нисходящим. Есть также ряд с установившимся трендом: TOTAL_CPD. Данные ряды имеют выраженную недельную сезонность. Компонента остатков не совсем похожа на случайный шум, хотя в идеале должно быть так.

Model: ARIMA

Stationarity test: проверим ряды на стационарность, используя тест Дики-Фуллера. **Нулевая гипотеза** H_0: $g = 0$ (существует единичный корень, ряд нестационарный). **Альтернативная гипотеза** H_1: $g < 0$ (единичного корня нет, ряд стационарный). Отвергаем H_0 на *5%* уровне значимости, если Test Statistic < Critical value[5%] => единичных корней нет и ряд стационарен в противном случае (H_0 верна) необходимо продифференцировать ряд и сделать проверку заново. Исходя из результатов теста будем осуществлять подбор параметра d.

Test for Russia:

```
Test: RUSSIA_CPD
Results of Dickey-Fuller Test:
Test Statistic                   1.447087
p-value                          0.997318
#Lags Used                      11.000000
Number of Observations Used    103.000000
Critical Value (1%)             -3.495493
Critical Value (5%)             -2.890037
Critical Value (10%)            -2.581971
dtype: float64
```

```
Test: RUSSIA_FPD
Results of Dickey-Fuller Test:
Test Statistic                   0.572733
p-value                          0.986924
#Lags Used                      13.000000
Number of Observations Used    101.000000
Critical Value (1%)             -3.496818
Critical Value (5%)             -2.890611
Critical Value (10%)            -2.582277
dtype: float64
```

```
Test: RUSSIA_C
Results of Dickey-Fuller Test:
Test Statistic                  -4.260927
p-value                          0.000519
#Lags Used                      13.000000
Number of Observations Used    101.000000
Critical Value (1%)             -3.496818
Critical Value (5%)             -2.890611
Critical Value (10%)            -2.582277
dtype: float64
```

```
Test: RUSSIA_F
Results of Dickey-Fuller Test:
Test Statistic                  -2.142223
p-value                          0.227879
#Lags Used                       9.000000
Number of Observations Used    105.000000
Critical Value (1%)             -3.494220
Critical Value (5%)             -2.889485
Critical Value (10%)            -2.581676
dtype: float64
```

Conclusion: таким образом для Russia, получим что ряды RUSSIA_CPD,RUSSIA_FPD,RUSSIA_F **являются стационарными** => d = 0, ряд RUSSIA_C является **не является стационарным** => необходимо продифференцировать (d!=0)

Test for Total:

```
Test: TOTAL_CPD
Results of Dickey-Fuller Test:
Test Statistic                  -0.955213
p-value                          0.769247
#Lags Used                      11.000000
Number of Observations Used    103.000000
Critical Value (1%)             -3.495493
Critical Value (5%)             -2.890037
Critical Value (10%)            -2.581971
dtype: float64
```

```
Test: TOTAL_FPD
Results of Dickey-Fuller Test:
Test Statistic                  -1.168983
p-value                          0.686794
#Lags Used                       8.000000
Number of Observations Used    106.000000
Critical Value (1%)             -3.493602
Critical Value (5%)             -2.889217
Critical Value (10%)            -2.581533
dtype: float64
```

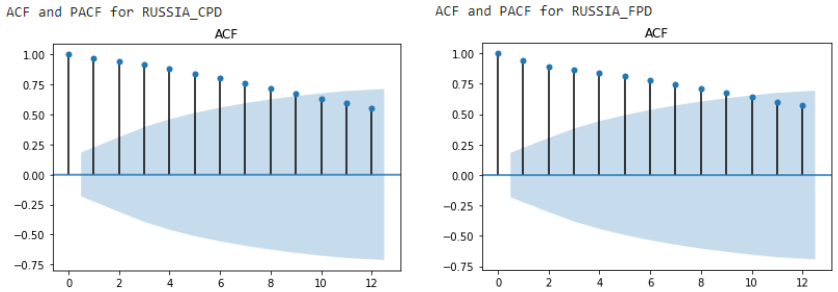
```
Test: TOTAL_C
Results of Dickey-Fuller Test:
Test Statistic                   1.198526
p-value                          0.995976
#Lags Used                      12.000000
Number of Observations Used    102.000000
Critical Value (1%)             -3.496149
Critical Value (5%)             -2.890321
Critical Value (10%)            -2.582122
dtype: float64
```

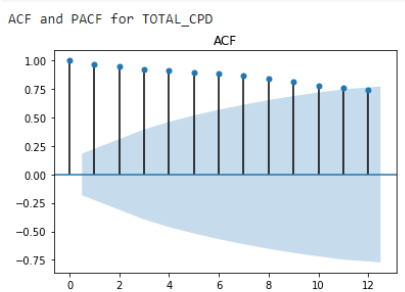
```
Test: TOTAL_F
Results of Dickey-Fuller Test:
Test Statistic                  -0.386553
p-value                          0.912320
#Lags Used                       9.000000
Number of Observations Used    105.000000
Critical Value (1%)             -3.494220
Critical Value (5%)             -2.889485
Critical Value (10%)            -2.581676
dtype: float64
```

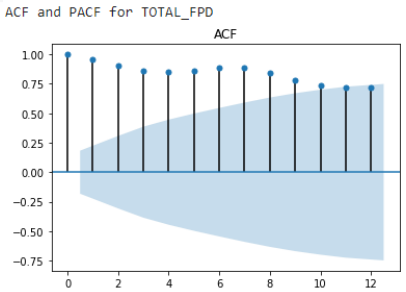
Conclusion: таким образом для Total, получим что ряды TOTAL_CPD,TOTAL_FPD,TOTAL_*F, TOTAL*_C **являются стационарными** => d = 0 и тогда модель будет ARMA

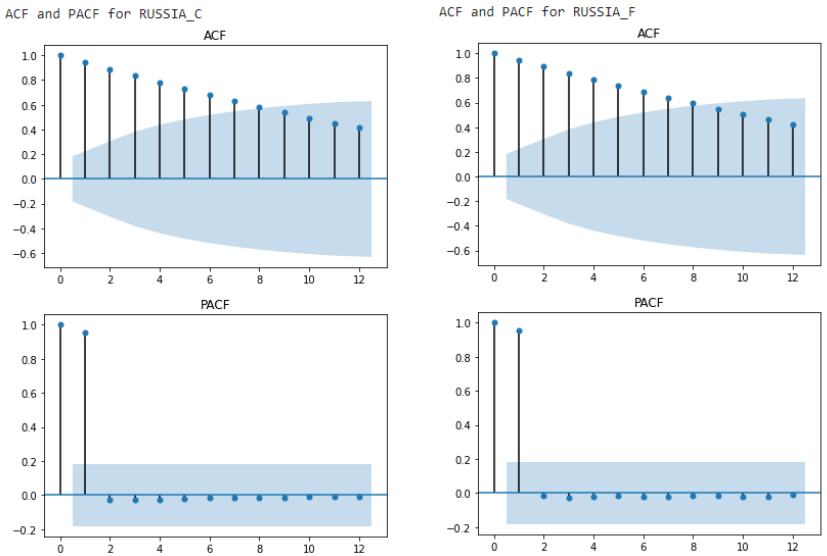
Model: ARIMA

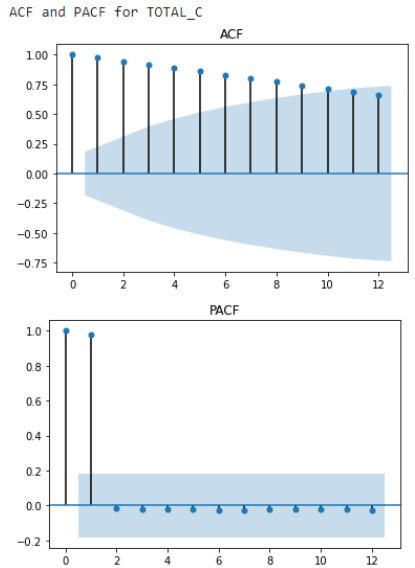
ACF and PACF plots: Построим графики автокорреляции, частичной корреляции из которых будем определять параметры p and q как количество сильно выступающих лагов на соответствующих графиках. Для графиков Russia значения p and q будут следующими: p = 9, q = 2; для Total аналогично. Исключения только: TOTAL_FPD and TOTAL_CPD, где значение q =4.

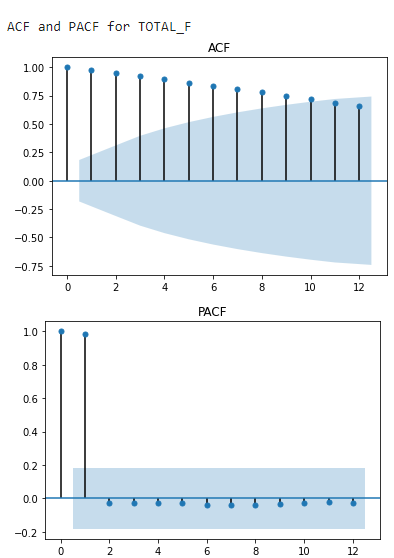












4) Making predictions.
Model: Prophet

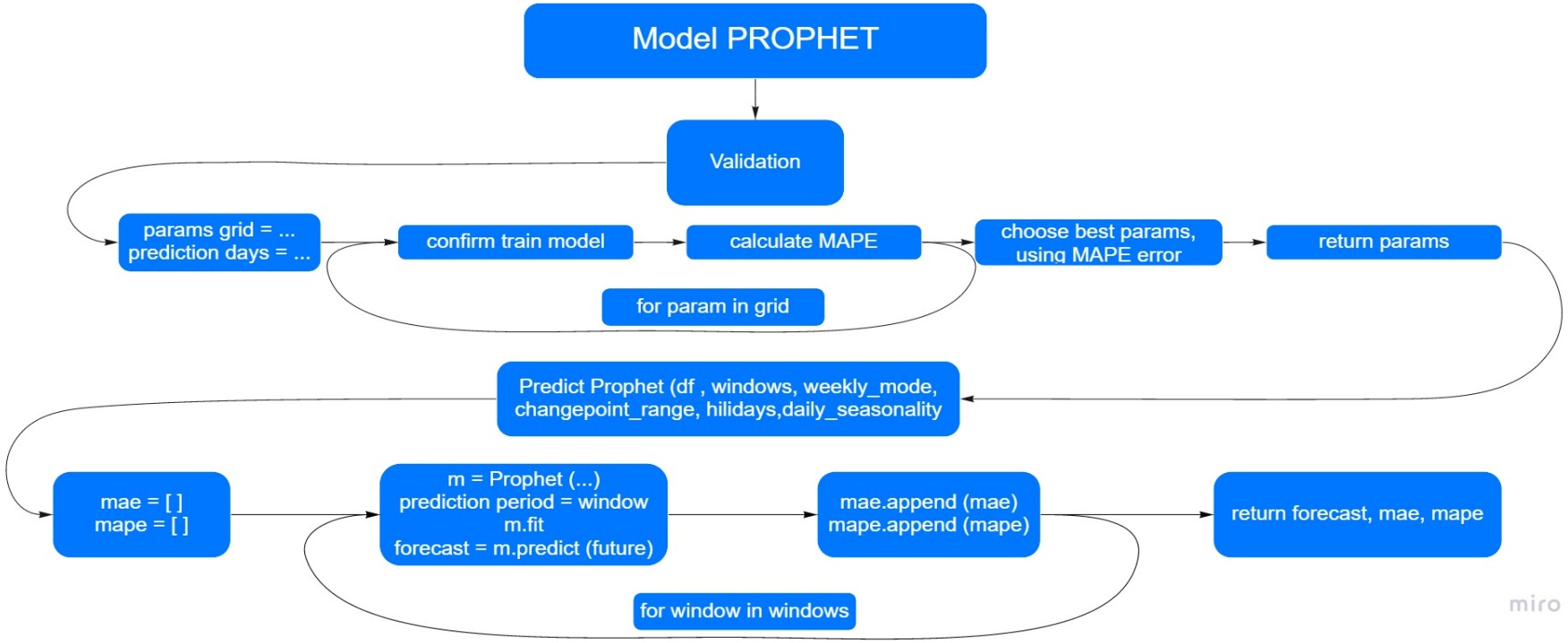
Model: Prophet

Концепция модели довольно проста. Она учитывает 3 компоненты: сезонность, тренд, аномальные дни и ошибку прогноза. Данная модель не требует стационарности от временных рядов, нужно лишь подобрать гиперпараметры таким образом, чтобы качество предсказаний было наилучшим. В документации сказано, что Prophet лучше всего работает с временными рядами, которые имеют сильные сезонные эффекты и несколько сезонов исторических данных и что Prophet устойчив к изменениям в тренде и отсутствующим и, как правило, хорошо справляется с выбросами. Проверим данные заявления на практике.

Перед построением нашей модели выполним валидацию для подбора следующих гиперпараметров:

а) seasonality mode – вид сезонности б) changepoint range – точки перелома

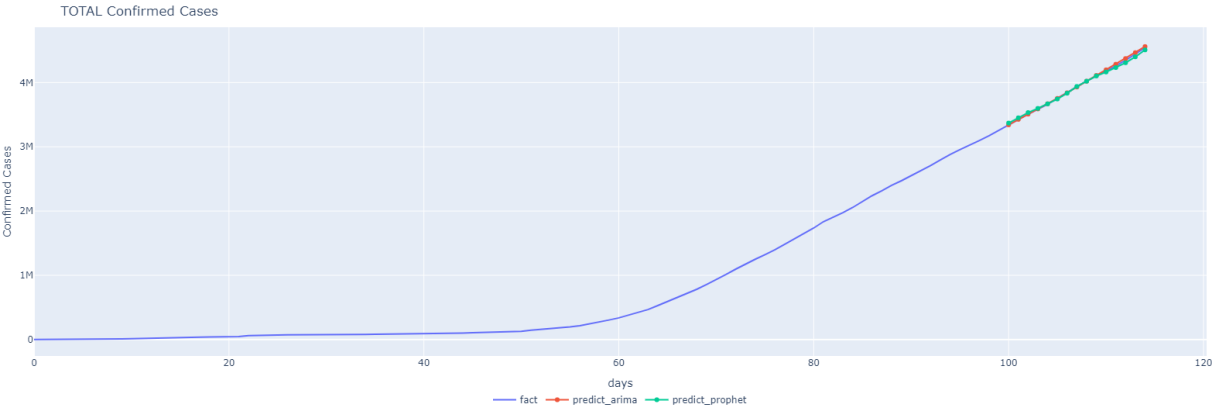
для улучшения качества прогнозов стоило подбирать и другие параметры модели, однако времени до дедлайна мало, поэтому остановимся на этих двух :) После чего выберем наилучшие гиперпараметры и используем их в модели.

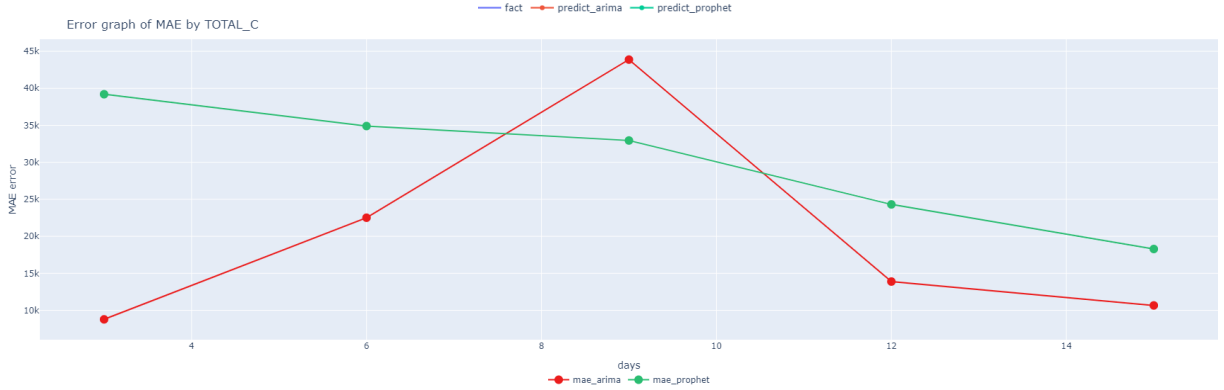


5) Visualize predictions and compare of model by metrics

Выполним визуализацию полученных результатов. Слева представлены графики для tolat, справа для Russia. Первые две графика – результат предсказания и действительный результат на 15 дней и действительный результат. Третий, четвертый график – сравнение ошибок MAE для двух моделей за 3,6,9,12 и 15 дней соответственно. Пятый, шестой график - сравнение ошибок MAPE для двух моделей за 3,6,9,12 и 15 дней соответственно.

Chart of prediction total confirmed cases and MAE, MAPE errors:





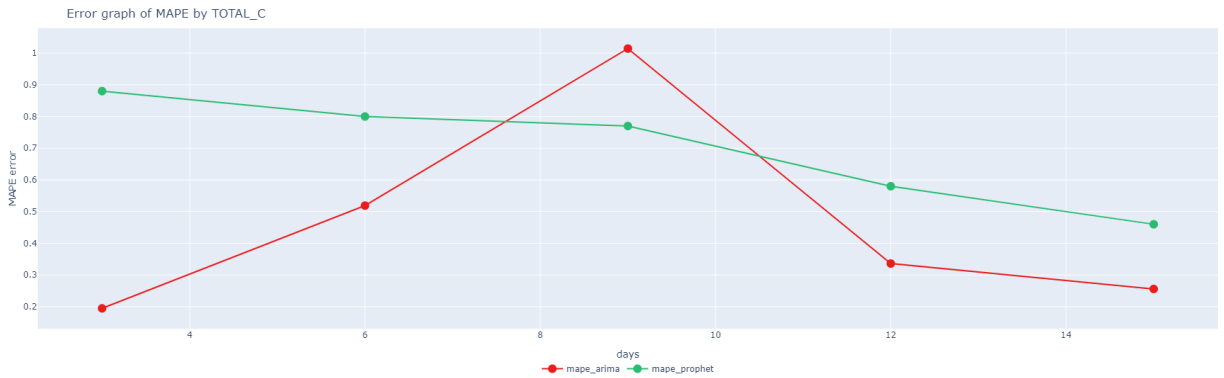
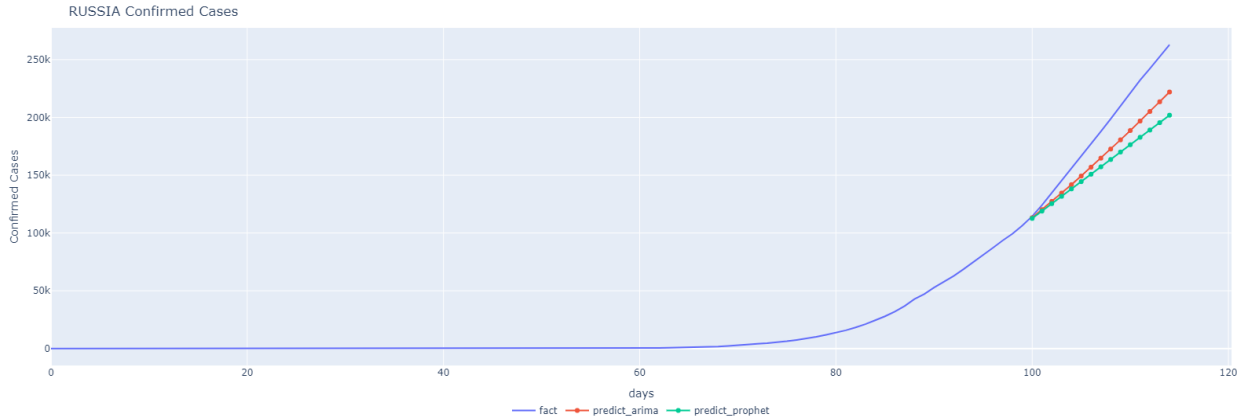
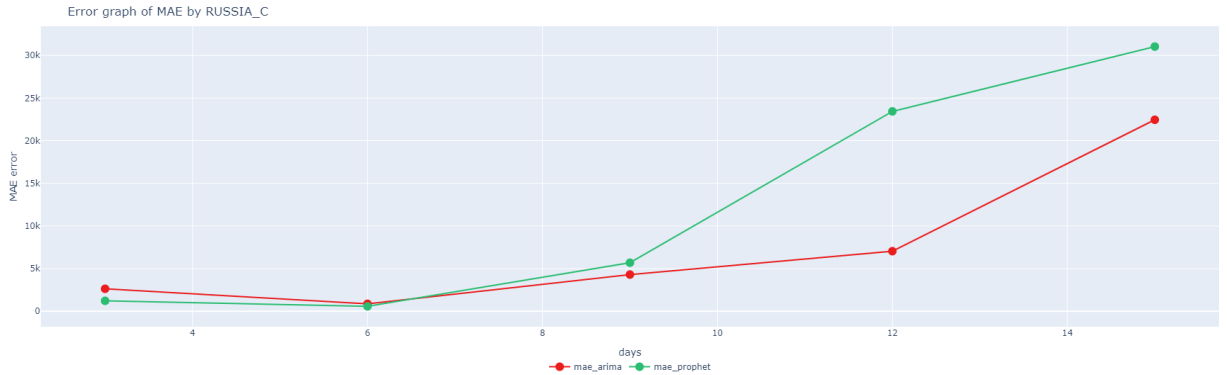


Chart of prediction Russia confirmed cases and MAE, MAPE errors:





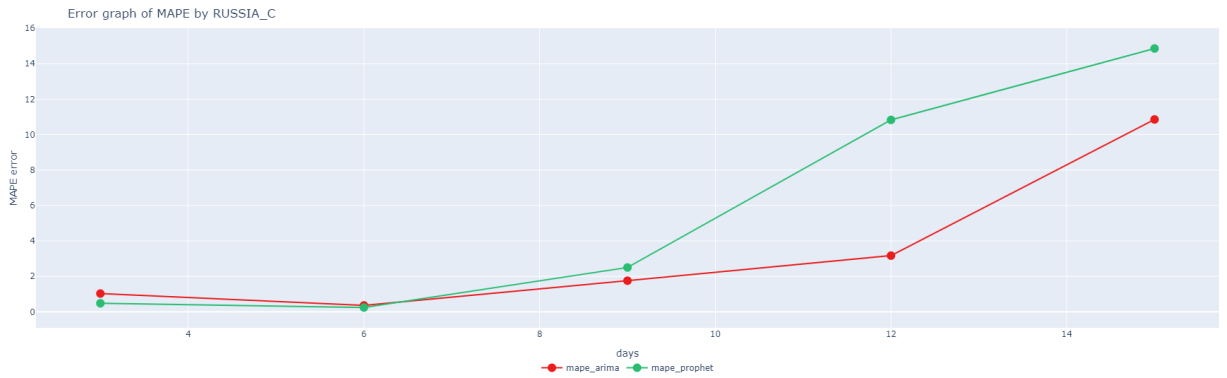
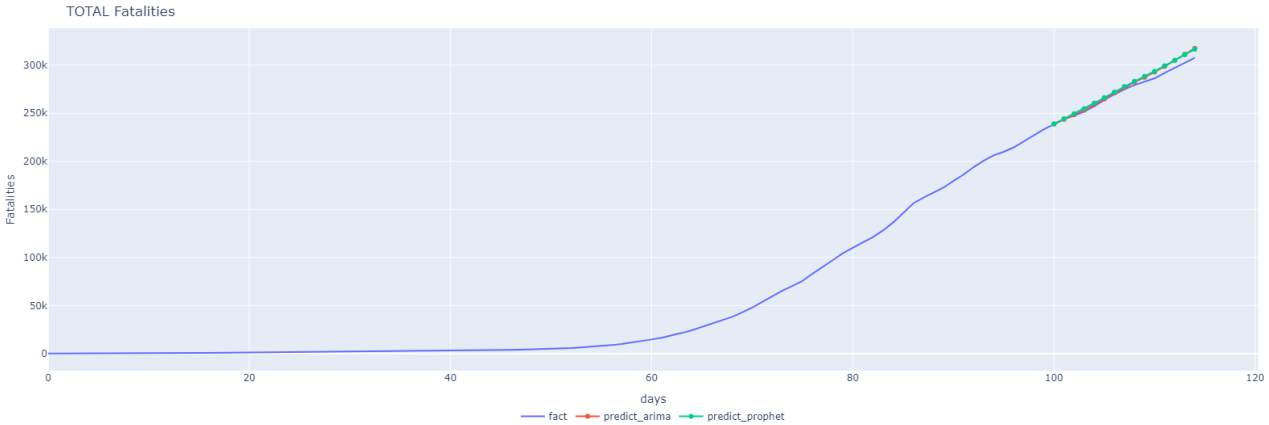
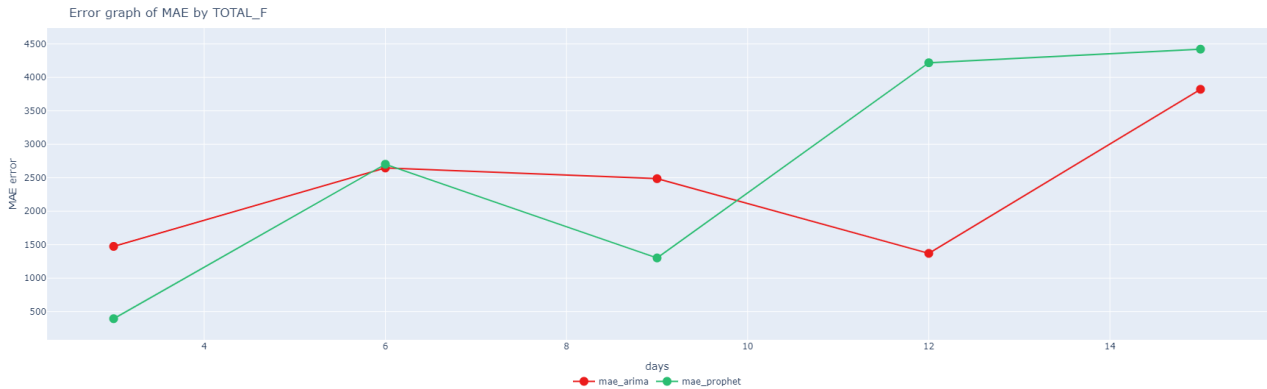


Chart of prediction total fatalities and MAE, MAPE errors:





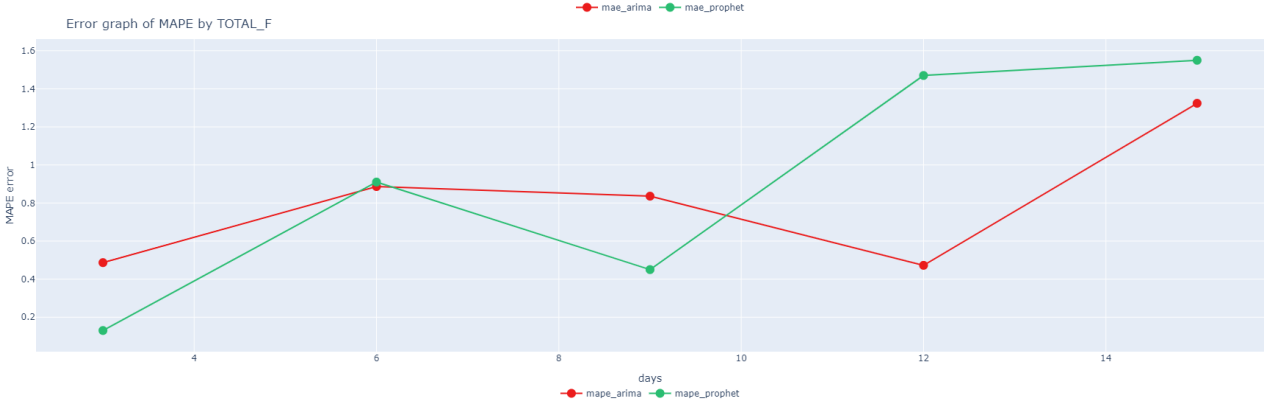
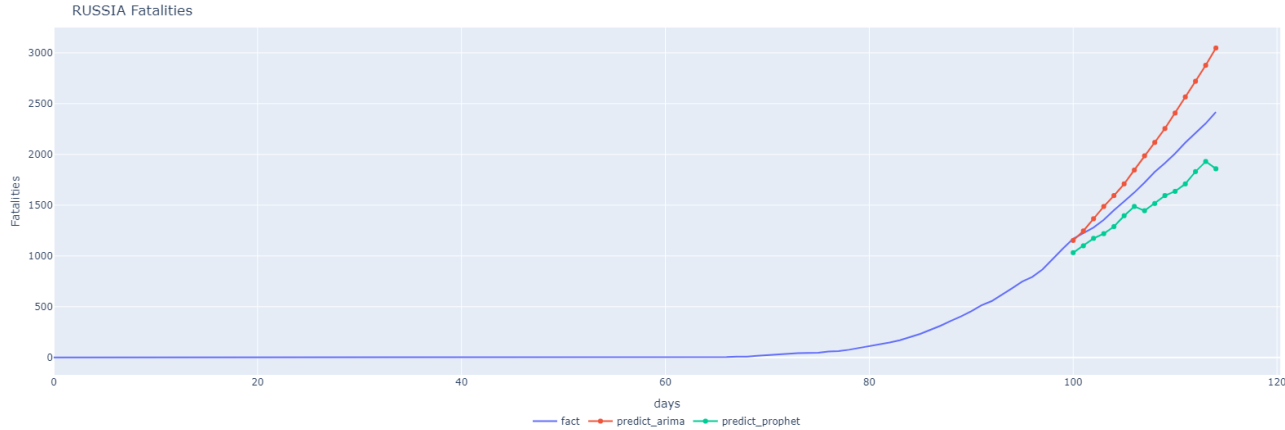
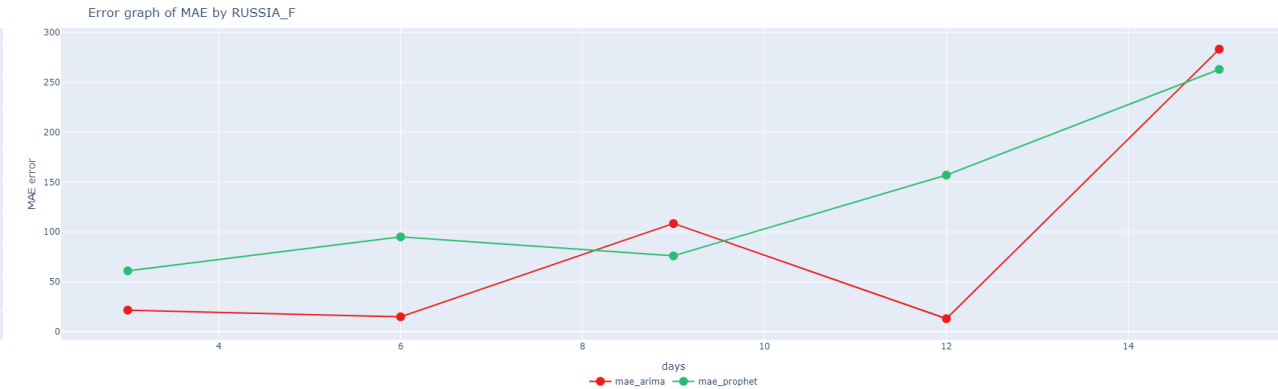


Chart of prediction Russia fatalities and MAE, MAPE errors :





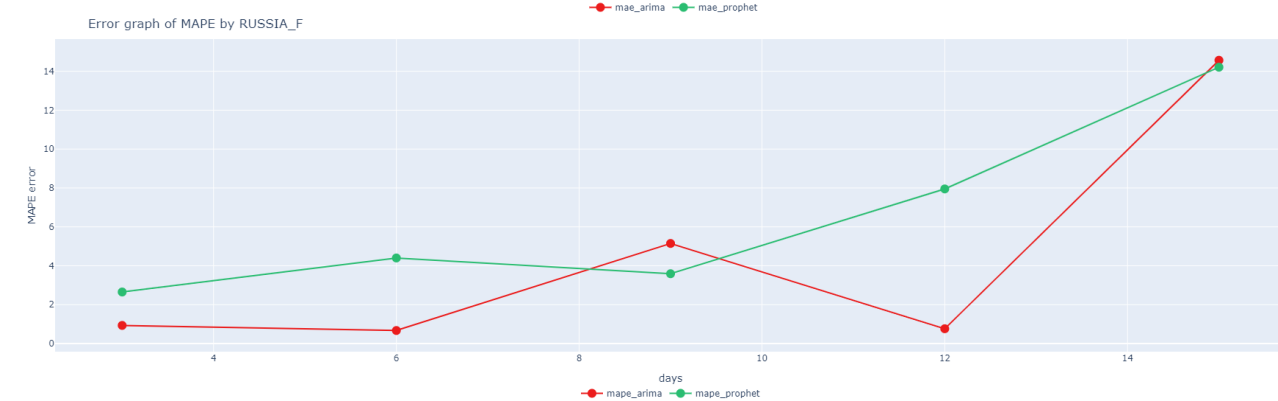
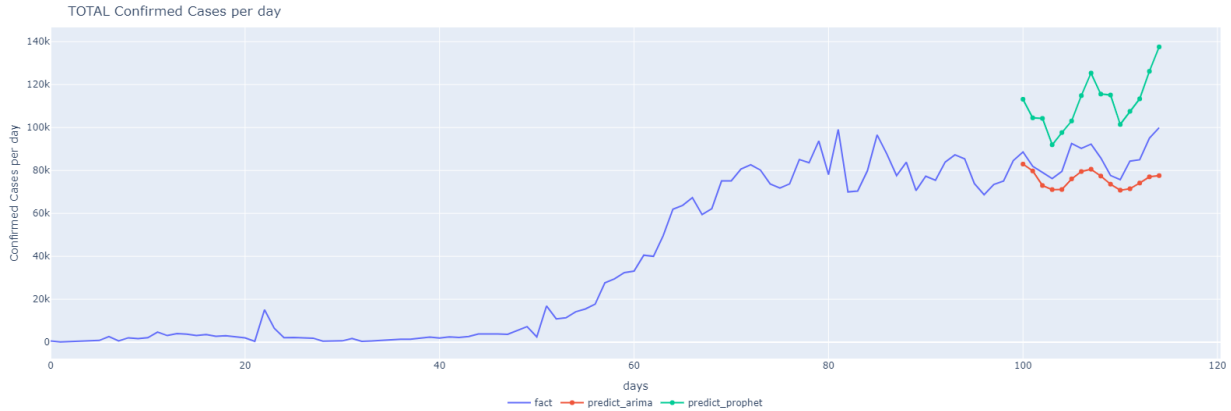
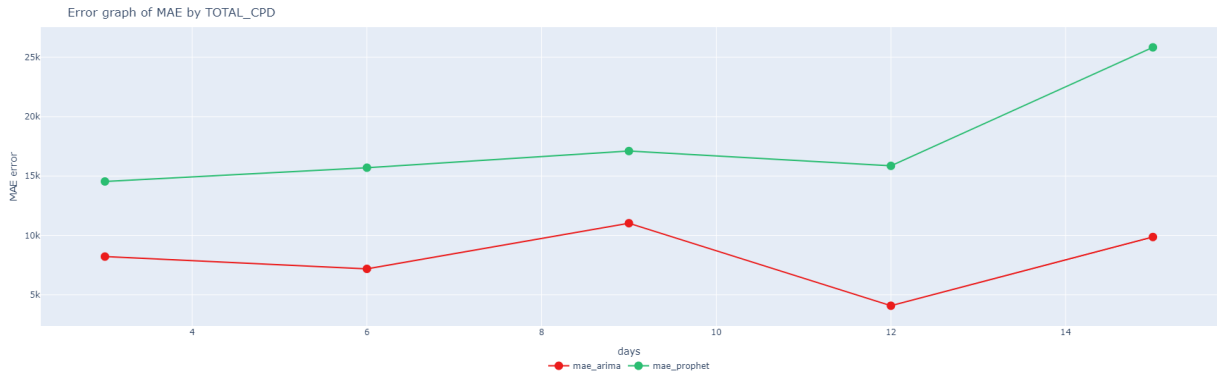


Chart of prediction total Confirmed Cases per day and MAE, MAPE errors:





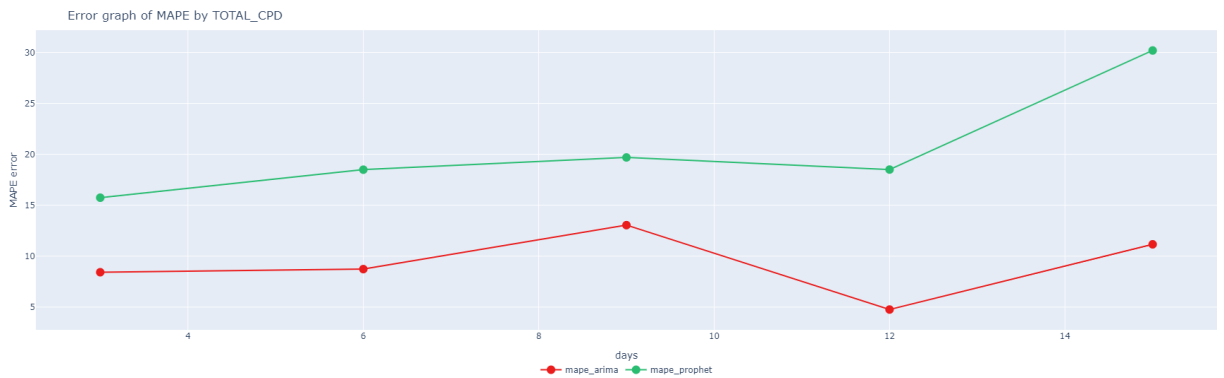
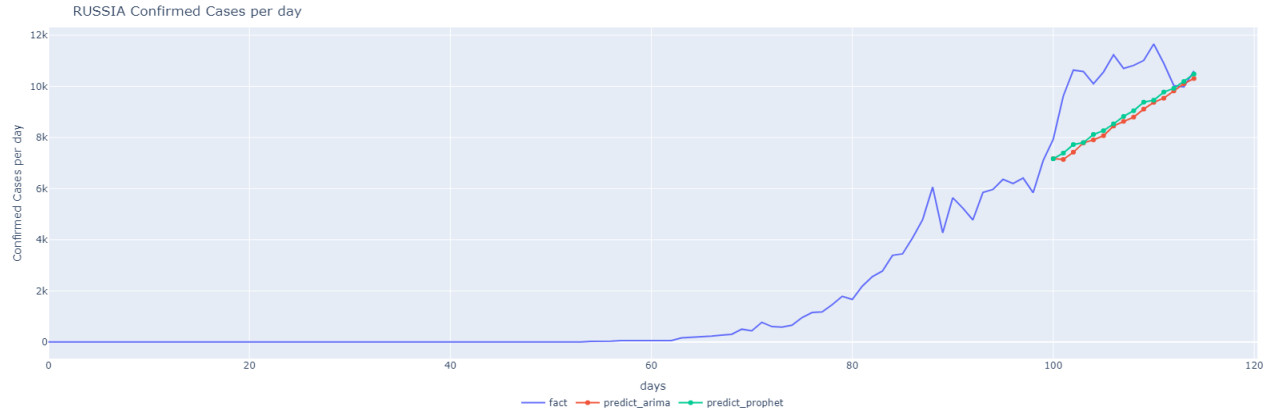
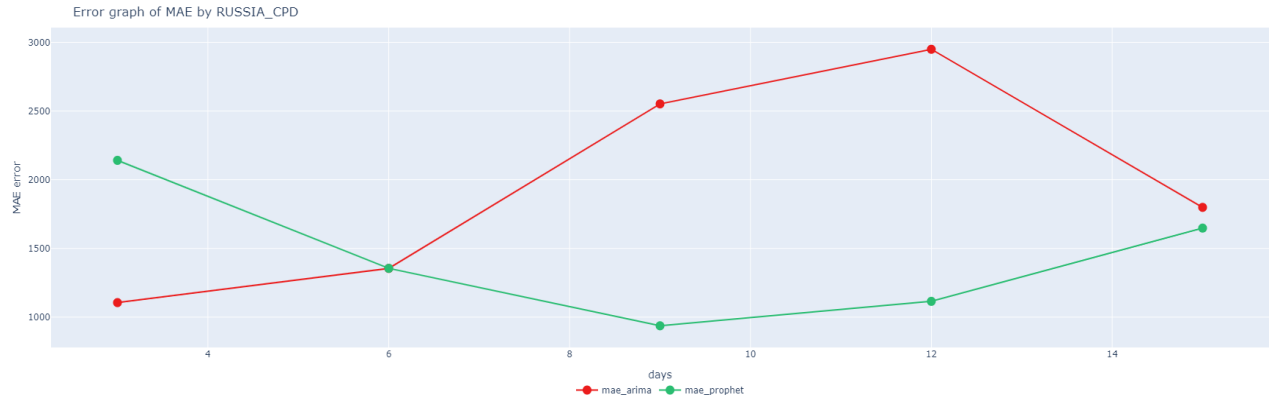


Chart of prediction Russia Confirmed Cases per day and MAE, MAPE errors:





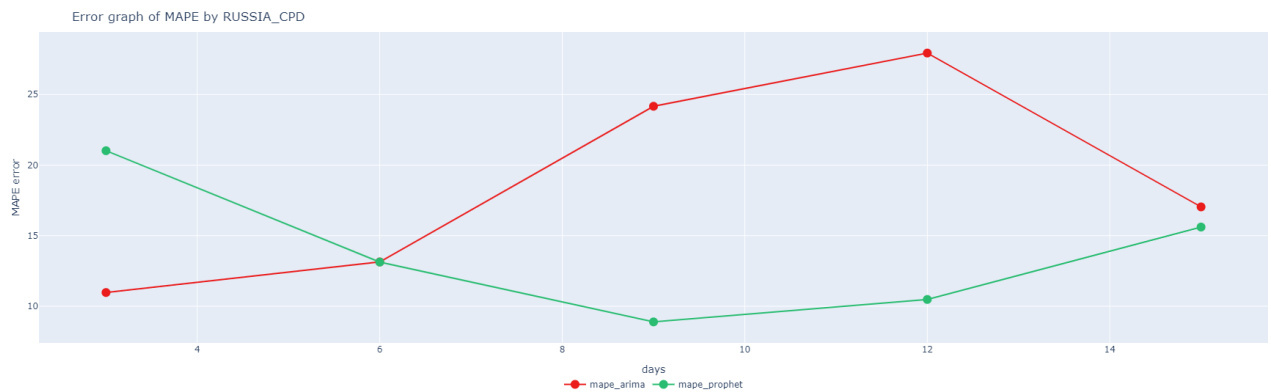
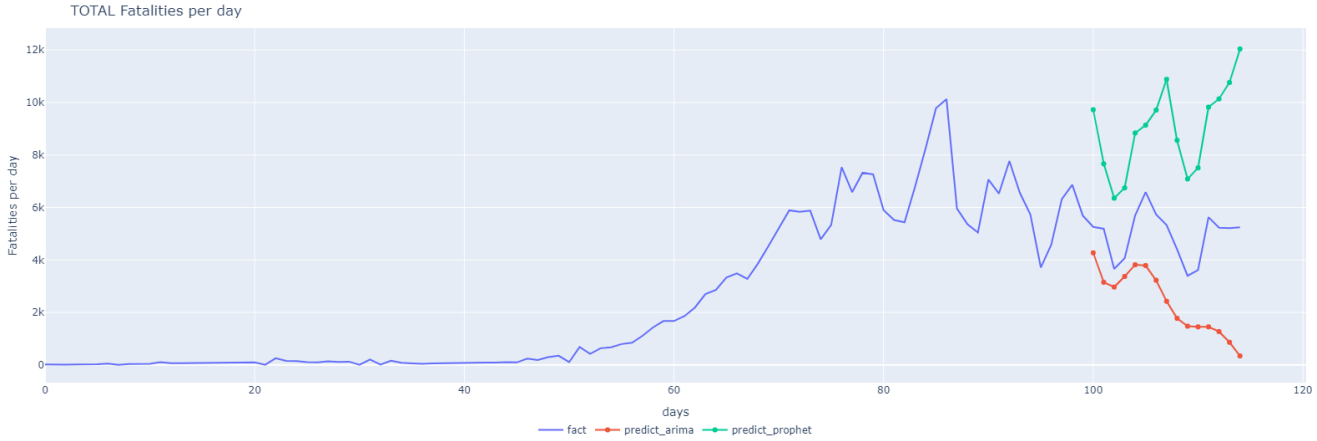
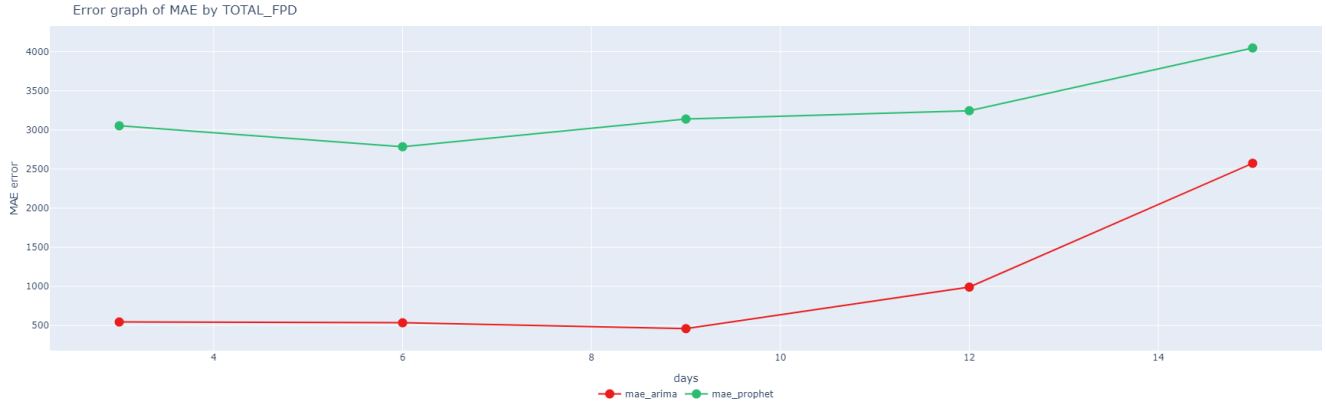
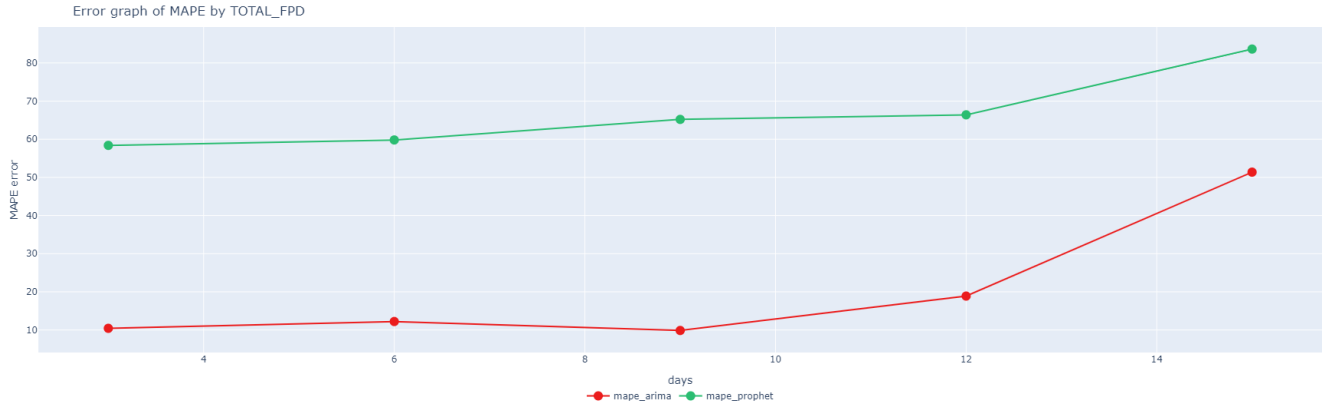


Chart of prediction total fatalities per day and MAE, MAPE errors:







Conclusion:

Проанализировав графики, можно сделать следующие выводы о моделях:

- При прогнозировании рядов, которые имели приближенную линейную зависимость (TOTAL_C, RUSSIA_C), обе модели справились довольно хорошо. Ошибка MAPE для TOTAL_C не превышала 1% при прогнозе до 15 дней. Однако при прогнозировании RUSSIA_C значение ошибки доходило до 15% у модели Prophet. Ошибка MAE для TOTAL_C достигала максимума у модели Arima и составила ~45к, т.е. в среднем модель ошибается на ~45к человек из ~4млн. Для TOTAL_C максимум MAE был у модели Prophet и составил ~30к. Стоит отметить, что для данных двух временных рядов модель предсказания Arima справилась с задачей чуть лучше.
- Аналогичные выводы можно сделать для TOTAL_F, RUSSIA _F (эти графики имеют линейную зависимость). Ошибки на feature TOTAL_F не были высоки (MAPE не больше 2%, при MAE ~4.5к), на признаке RUSSIA _F MAPE и MAE, как ожидалось, оказались выше (MAPE ~14%, MAE ~ 250). В данном случае обе модели имели примерно одинаковые ошибки на горизонте 15 дней, из чего можно сделать вывод о том, что при прогнозировании TOTAL_F and RUSSIA _F модели оказались равными.
- При прогнозировании рядов, не имеющих выраженных линейную зависимость (TOTAL_CPD, RUSSIA_CPD ,TOTAL_FPD), predict моделей ожидаемо оказался хуже. Так, модель Prophet TOTAL_CPD на 15 дневном прогнозе завышала результаты, а модель Arima, наоборот, занижала (max MAPE Prophet ~ 30%, a max MAPE Arima ~14%). В данном случае для TOTAL_CPD предсказания Arimы были лучше. Однако при предсказания ряда для RUSSIA_CPD модель Prophet справилась лучше (max MAPE Prophet ~ 15%, a max MAPE Arima ~28%)
- Также хочу отметить тот факт, что для прогноза было взяты данные всего за 4 месяца, что могло привести к недообучаемости у моделей

6) Conclusions

Conclusion to Project:

Итак, сделаем выводы по проделанной работе и ответим на следующие вопросы:

1) Какая модель предсказаний оказалась наилучшей ?
Как можно было заметить из предыдущего слайда, при разных временных рядах модели ведут себя по разному, поэтому однозначного ответа на данный вопрос дать нельзя. Тут необходимо рассматривать конкретный случай и делать выводы по нему. Можно сказать одно: обе модели неплохо предсказывают линейные зависимости, но довольно сильно ошибаются, если в данных появляется нелинейный тренд.

2) При предсказании какого временного ряда были достигнуты наименьшие/наибольшие ошибки ?
Самые минимальные ошибки MAE , MAPE были при прогнозировании временного ряда TOTAL_C и составили: MAPE ~1%, MAE ~ 25к.
Максимальные ошибки MAE , MAPE были при прогнозировании временного ряда TOTAL_FPD и составили: MAPE ~70%, MAE ~ 4к.
Как можно было заметить низкое MAPE оказалось при довольно высоком MAE и наоборот: высокое MAPE при низком MAE . Поэтому необходимо рассматривать метрики в совокупности, чтобы правильно интерпретировать результаты.

3) Что влияет на качество предсказаний, можно ли улучшить точность предсказаний моделей ?
Ошибка моделей зависит от многих факторов: распределение данных, количество дней прогноза и тд.
Качество обоих моделей можно улучшать применяя различные методы. Например, кросс-валидацию вместо обычной. Также можно было использовать большее количество параметров в grid search для модели Prophet, чтобы найти самые наилучшие. Тогда качество прогноза должно стать ещё лучше.

4) Нужно ли носить маски и перчатки, чтобы не заразиться COVID 19 ?
Ответ однозначный: Да, конечно! ☺

Спасибо за внимание!

Контакты для связи:
-почта: vitya.zapselsky@Yandex.ru
-телефон: +7-(920)-511-51-21