Итоговое задание: Анализ данных с помощью инструментов Excel

Асем Кусаинова, Нуркен Адилбек

В данной работе был проведен анализ набора данных по ежедневной заболеваемости COVID-19 среди стран всего мира за 2020 год. Датасет COVID-19-geographic-distribution содержит в себе 50203 записей и 12 атрибутов, включающих в себя дату, составляющие которой также разделены на отдельные поля (день, месяц, год), количество заболеваний и смертей, названия стран с их территориальными кодами и размером населения за 2019 год, принадлежность к континенту и совокупное количество заболеваний COVID-19 за 14 дней на каждые 100000 человек.

На первоначальном этапе была проведена оценка качества данных, на предмет различных аномалий, наличия пропущенных значений, ошибок и не соответствующей кодировки. В датасете было выявлено отсутствие статистики по месяцам ноябрь и декабрь 2020-го года. Кроме того, в полях по случаям заболевания и смертям среди населения по COVID-19 присутствуют отрицательные значения, общее количество записей которых составляет 25 строк. Так как данное количество составляет малую долю всех данных, рекомендовано исключить указанные записи при дальнейшем построении прогнозных моделей. В атрибуте имен стран была обнаружена ошибка в написании названия страны CuraÃ§ao, и была изменена на Curaçao. В дополнении, атрибуты по popData2019 в стране Cases on an international conveyance Japan и Cumulative number for 14 days of COVID-19 cases per 100000 для всего датасета содержат большое количество пропущенных значений, что может повлиять на прогнозирование потенциальной модели.

Следующим шагом было понять слабые и сильные стороны представленных данных, определить их достаточность, предложить идеи, как их использовать, для чего были использованы графики и статистические показатели. С помощью визуализации уровня заболеваемости и смертности по имеющимся атрибутам были получены следующие наблюдения. Согласно имеющимся данным по распределению количества заболеваний по континентам (см. График 1), самые ранние случаи заболевания зафиксированы в Китае в декабре 2019, а наибольшее количество заболеваний и смертей за весь период произошли среди населения континента Америка.

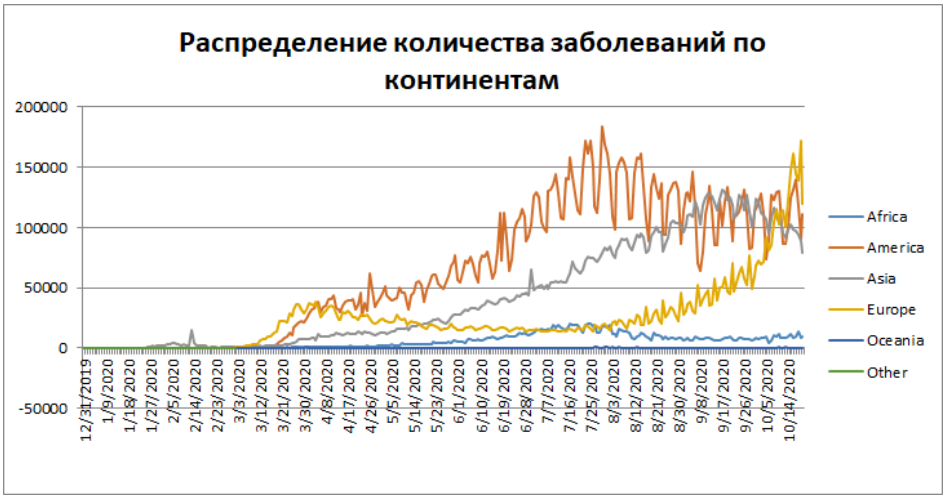

График 1.

По графику общего числа заболеваний, наибольшее количество случаев COVID-19 было в августе 2020 года со значением приближенным к 8.5 миллионам людей. Также, показатели по заболеваниям и смертям в начале года имеют экспоненциальный рост с последующей стабилизацией темпа ближе к середине года. Касательно доли смертности среди всех случаев заболеваний, при нахождении разницы между общим количеством заболевших и летальных исходов, определяется статистика количества выздоровевших и тех, кто на стадии выздоровления. Исходя из этих вычислений, процент выздоровевших и смертей составил 97,24% и 2,76% соответственно, что соответствует реальным цифрам указанных в статьях о COVID-19. Кроме того, во время проведения анализа данных, было получено значение корреляции между количеством заболеваний и смертей составляющее 0,73, что создает предположение о росте смертности с увеличением случаев заболевания COVID-19 и наоборот, отсутствие смертей при низком уровне заболеваний.

На основе полученной информации, была рассмотрена сходимость количества заболеваний, смертей и совокупное количество заболеваний COVID-19 за 14 дней на каждые 100000 человек под доступные

линии тренда, методы регрессии, а также прогнозирование с помощью временных рядов. Все предсказания по показателям основаны на исторических данных. Было выявлено, что значения поля заболеваний соответствуют полиномиальной линии тренда 3 степени с коэффициентом детерминации равным 0.9619. При построении листа прогноза основанном на экспоненциальном сглаживании предсказания на различные промежутки времени имеют относительно небольшую ошибку при сравнении с реальными данными. Однако, полученные данные не могут быть обобщены и использованы в полноценной мере в реалии ввиду отсутствия дополнительных данных по влияющим факторам на уровни заболеваемости и смертности.

График 2.

В связи с вышеуказанной информацией, для данного датасета было предложено предсказание распространение вируса на основе темпа роста (dissemination rate) в качестве коэффициента для линейной зависимости количества заболеваний от количества прошедших дней с начала эпидемии. Для построения прогноза были рассчитаны данные по распространению вируса в течение каждых 5 дней. Значения dissemination rate были получены при расчете отношения последних 5 дней к предыдущим данным количества заболеваний по последним 5 дням. Для прогноза количества заболеваний и смертности были рассмотрены несколько сценариев. Исходя из графика (см. График 2) можно предположить, что к середине марта 2020 в связи с высоким темпом распространения болезни был введен всемирный локдаун с последующим ослаблением. При имеющихся действительных данных общее количество заболеваний составляет 40472505, из которых около 1 млн. смертей. Один из кейсов, при котором ни одна из стран не предприняла необходимые ограничительные меры (локдаун, запреты на передвижение и т.д.), предполагает, что темп роста оставался бы неизменным как самое большое значение в середине марта равное 1.75; следовательно, по прогнозам наблюдается 1,13287E+16 общих случаев заболеваний до конца октября 2020 года. В теории, к середине года всё человечество вошло бы в категорию заболевших, что привело бы к высокому уровню смертности. В другом случае, предположим, что во время первых проявлений коронавируса в декабре 2019 были приняты все возможные ограничительные меры во всех странах, что по прогнозу привело бы к сравнительно низкому dissemination rate равному 1.0579 к концу февраля, и при данном значении согласно прогнозу к концу осени 2020 общее число зараженных ограничивается в 1179170 человек, из которых летальный исход составляет 32 тысячи случаев.

В заключении, в ходе данной работы была проделана предобработка и анализ качества данных, визуализация и исследование присутствующих трендов для прогнозирования, а также построение модели прогноза при разных сценариях развития роста заболевания COVID-19. В целом, в датасете присутствуют небольшие ошибки и пропущенные значения атрибутов, а также значения не входящие в реальный диапазон полей количества заболеваний и смертей, что незначительно влияет на качество датасета. По имеющимся данным возможно изучить основные тренды и статистику течения эпидемии по всем странам мира за 2020 год, однако качественное прогнозирование требует дополнительного актуального на нынешний день объема данных по всевозможным влияющим факторам, так как предсказания выполненные в данной работе основаны только на одном поле без зависимостей с какими либо другими переменными. Предлагаемый метод прогноза с использованием рассчитанных данных темпа развития заболеваний теоретически учитывает влияние вводимых ограничительных мер, которые могут влиять на количество заболеваний в целом в сравнении с ограничением методов скользящее среднее и простой регрессии, которые основываются на самой таргет переменной.