Смертность в период пандемии коронавируса

Данный вопрос будет исследован с нескольких сторон: как динамика смертности внутри одной страны и как избыточная смертность среди набора стран. В первом случае исследование проводится на переменных, непосредственно связанных с коронавирусной инфекцией таких как количество новых тестов, новых вакцинированных людей и ограничительные меры государств. Во втором случае будет возможность рассмотреть избыточную смертность, на которую пандемия влияла как непосредственно, так и косвенно. Кроме того, в такой вид исследования включен более широкий спектр показателей, например, количество сердечно-сосудистых заболеваний и ВВП на душу населения.

Динамика смертности внутри одной страны

Для исследования динамики смертности во времени была выбрана Франция, так как по данной стране имеется самый полный набор данных, покрывающий весь период пандемии. Однако, в связи с ошибкой в датасете – проставлено 0 смертей каждый день в течение нескольких месяцев – исследуемый период сокращен до промежутка с 2020-05-20 по 2022-06-17. Тестовый период составляет 10% и начинается с 2022-03-31. Также в связи с тем, что качество моделей значительно падает при удалении схожих переменных, сохранены все переменные, что немного усложняет интерпретацию.

Сравнение качества моделей

Модель	R-score	MAPE
LightGBM Regressor	0.09	1.95
XGBoost Regressor	-0.03	0.64
LinearRegression	-0.07	2.69
AutoARIMA	-0.41	-
KNeighborsRegressor	-0.14	4.77
DecisionTreeRegressor	-0.07	-
BayesianRidge	0.26	0.52

Так как коэффициент детерминации означает долю дисперсии зависимой переменной, объясняемой моделью, а в данном случае важна интерпретируемость результата, он выбран как основной. Лучшей моделью оказалась байесовская линейная регрессия, скорее всего из-за маленького наборы наблюдений. Данная модель и LightGBM выбраны для дальнейшей интерпретации.

LightGBM Regressor

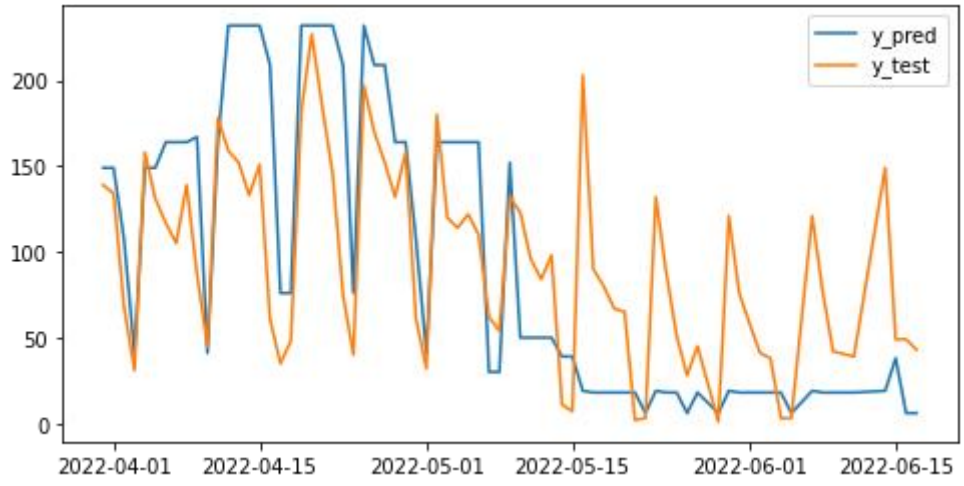

Рисунок 1. Значения, предсказанные моделью LightGBM Regressor.

Как видно, модель перестает повторять динамику данных после примерно полтора месяца. Это может быть связано как с неспособностью модели описать данные, так и с тем, что ситуация может значительно измениться за это время за счет большей доли вакцинированных и переболевших или изменения политики.

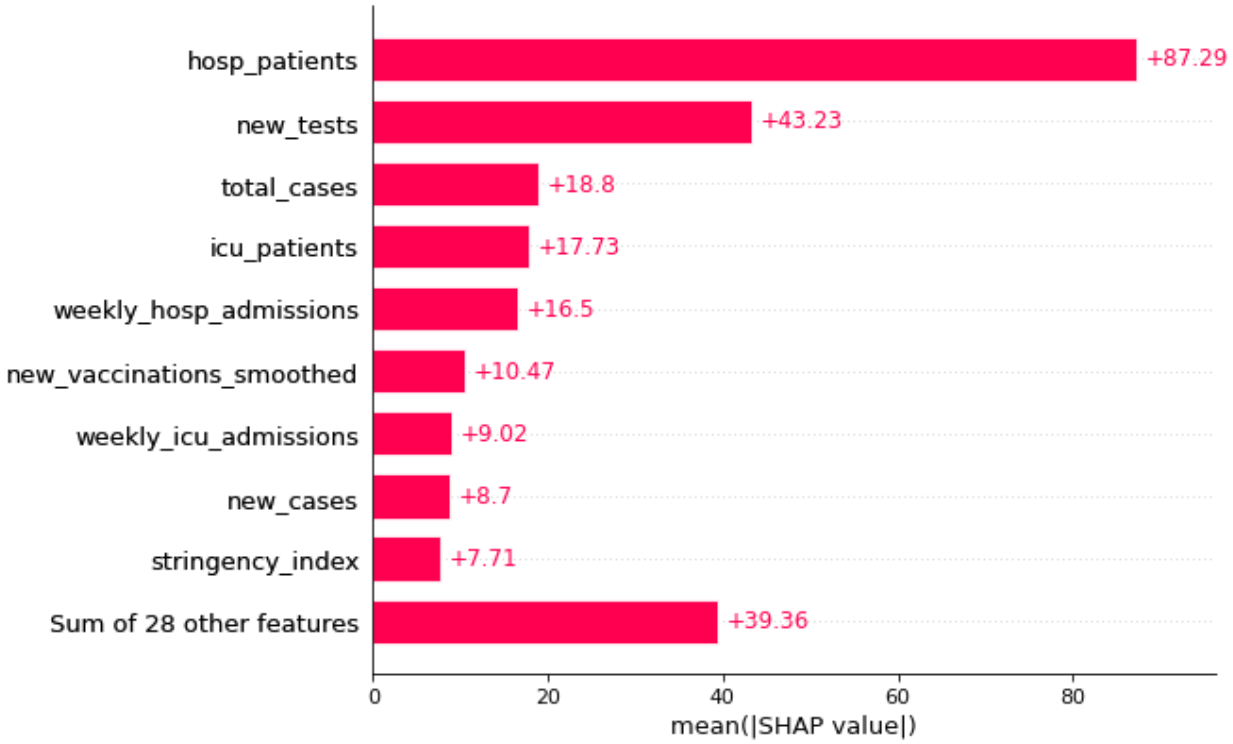

Рисунок 2. Важность переменных в соответствии со значением shaply.

Самыми важными оказались переменные, имеющие очевидную сильную связь со смертностью, например, общее количество случаев заражения и количество госпитализированных заболевших. Кроме того, наиболее значимыми являются показатели новых примененных вакцин и ограничения, установленные государством. Это возможно говорит о том, что смертность можно снижать вакцинированием, а ограничения, наложенные на население, имеют положительный эффект. Модель выделяет сглаженное количество новых прививок, скорее всего, в связи с тем, что это запаздывающая переменная, указывающая на количество вакцинированных ранее людей.

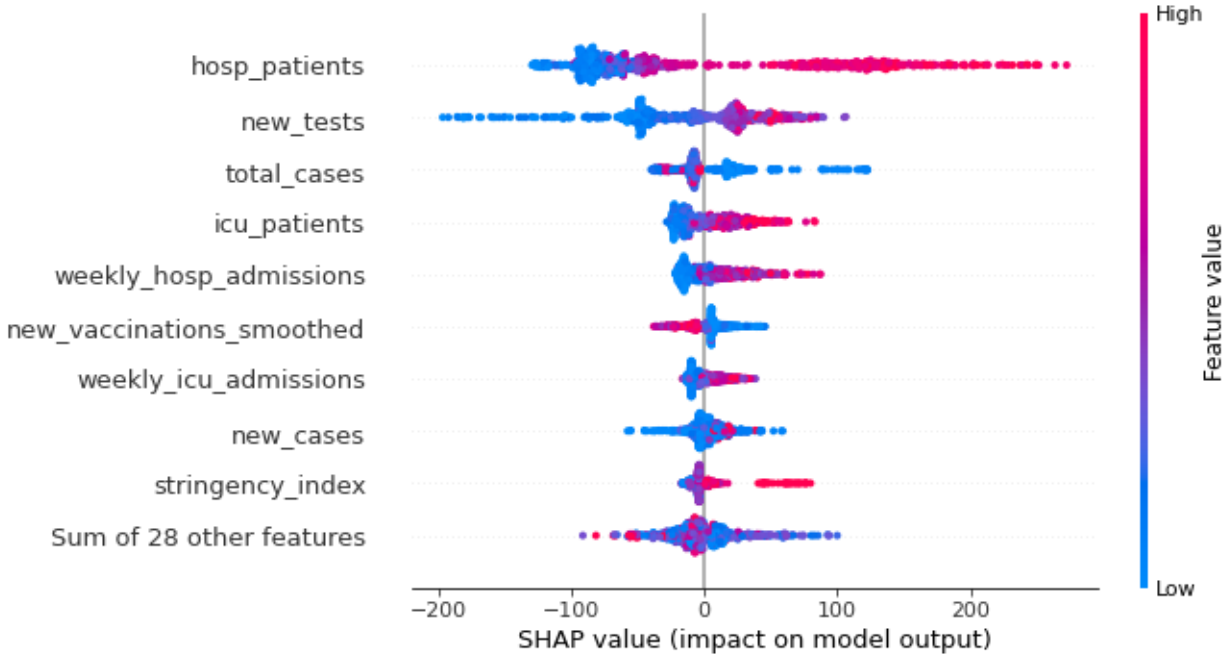

Рисунок 3. Значения shapley.

Рисунок 3 дает возможность лучше понять переменные, представляющий наибольший интерес – количество примененных вакцин и строгость ограничений, вводимых государством. Как видно на данном графике, количество вводимых ограничений имеет высокую значимость при положительном влиянии на большинство значений. Это говорит о том, что ограничения на население, согласно значениям shaply, не оказывают отрицательного влияния на текущую смертность, скорее являются реакцией на нее. Вакцинация тем не менее, скорее всего действительно может снижать смертность.

В подтверждении этому можно также взглянуть на Рисунок 3, на котором более подробно видно влияние самых значимых переменных на одно наблюдение.

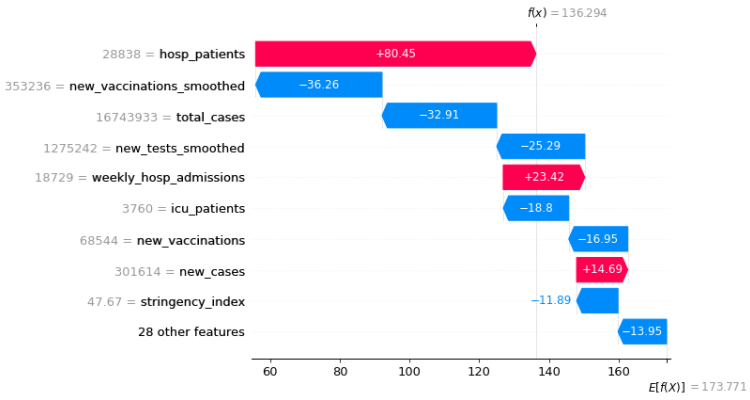

Рисунок 3. Значения shapley для одного наблюдения середине пандемии.

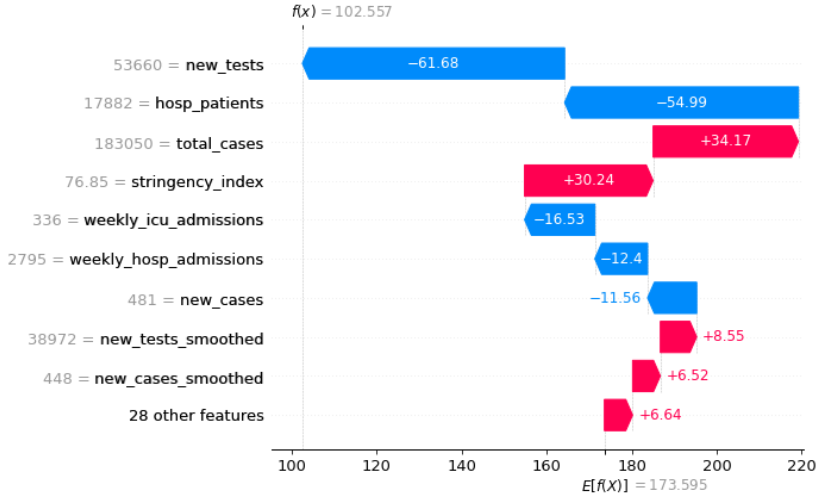

Рисунок 4. Значения shapley для одного наблюдения середине пандемии.

На Рисунке 4 видно, что до появления вакцины единственной непосредственно контролируемой обществом или государством значимой переменной было количество новых тестов, что, возможно помогало предотвращать тяжелые случаи болезни.

Bayesian Ridge

Данная модель имеет наиболее высокое качество, но менее интересную интерпретацию. Наиболее значимыми значениями обладают переменные, связанные с вакцинацией, что может только подчеркнуть её важность.

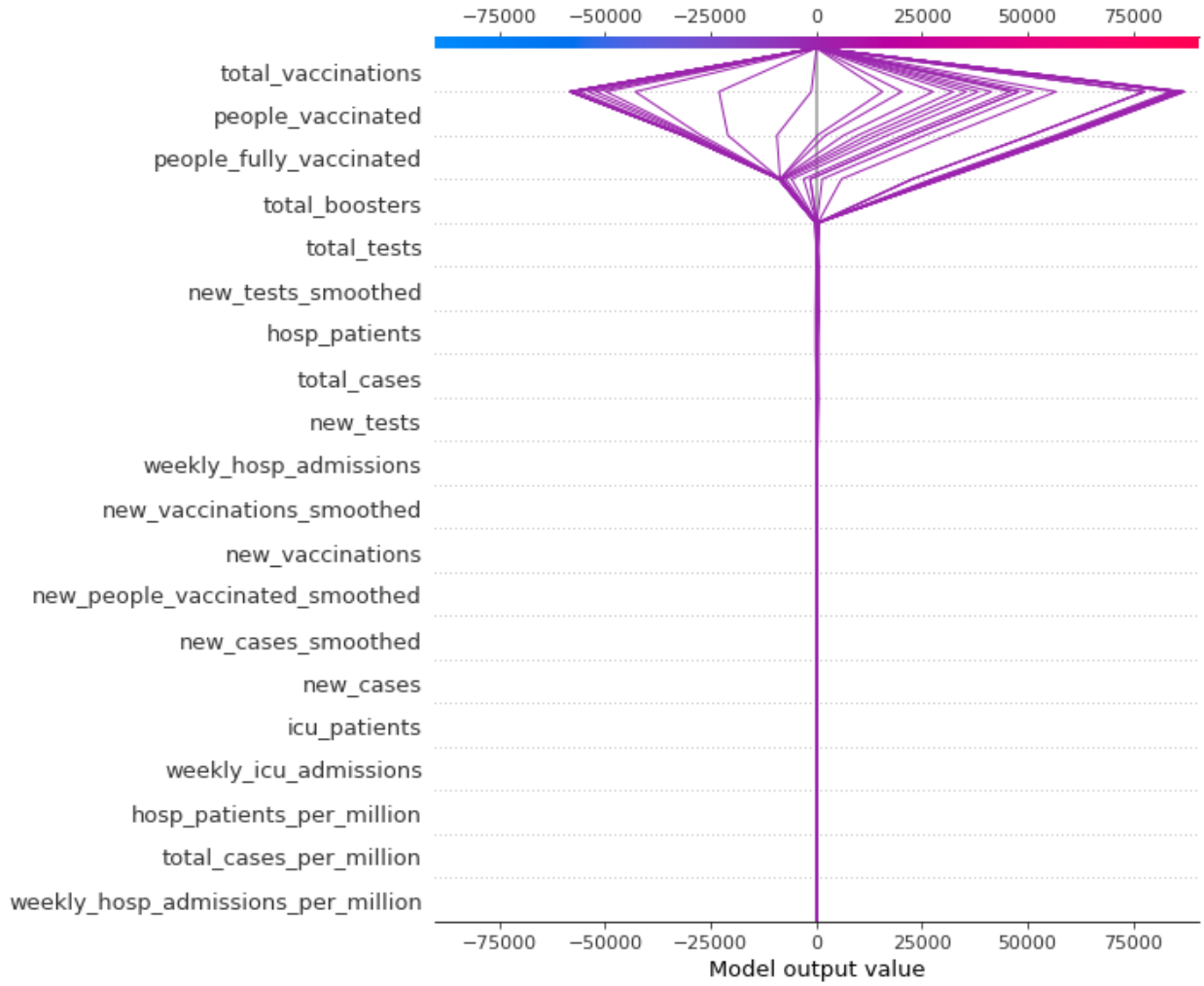

Рисунок 5. Визуализация значений shapley для модели байесовской линейной регрессии.

Избыточная смертность

Всего в выборке имеется 38 стран, большинство из которых находится в Европе в связи с большей информационной прозрачностью. Тем не менее созданные дамми-переменные, отвечающие за континент, в котором располагается страна, оказывались незначимыми в регрессиях.

Из переменных с ежедневной частотой созданы новые переменные, означающие средний показатель за весь период. В конечной модели остались две таки переменные:

1. avg_icu_patients_per_million – среднее количество пациентов в интенсивной терапии на миллион человек;
2. avg_hosp_patients_per_million – среднее количество госпитализированных пациентов на миллион человек.

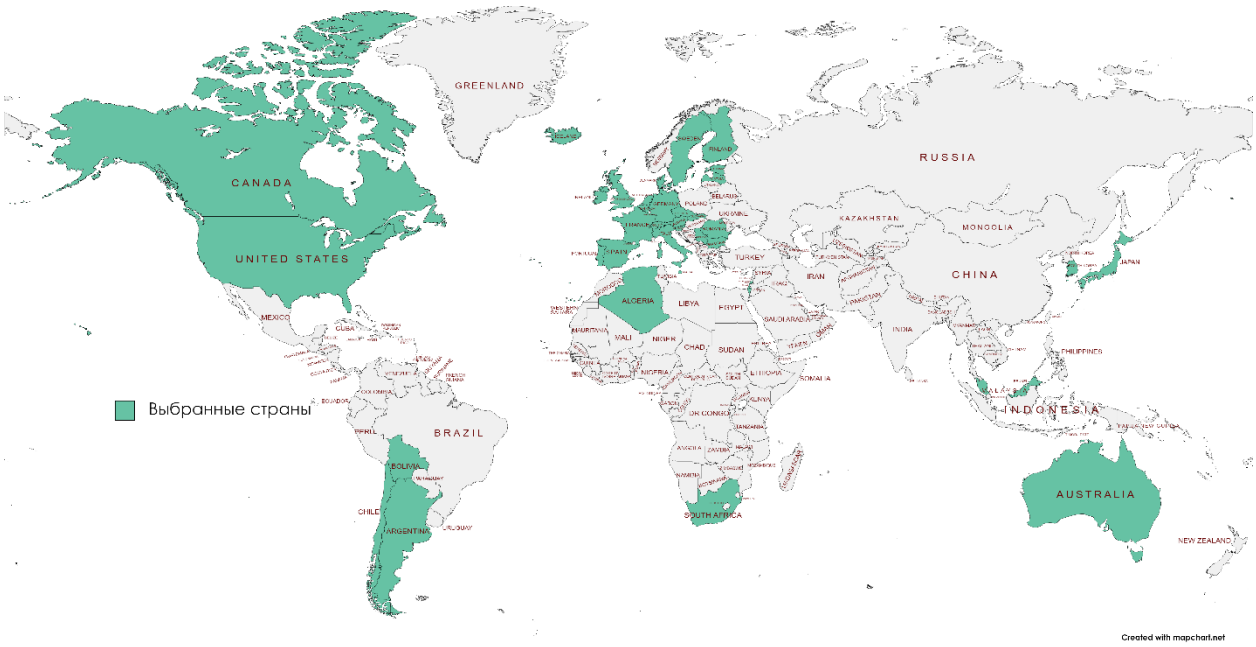

Рисунок 1. Визуализация выборки стран.

Так как в данном случае важна интерпретируемость модели оставим только значимые и не мультиколлинерные переменные. Мультиколлинеарные переменные могут способствовать нестабильности коэффициентов и увеличивать значимость переменных. Также остатки выбранной модели прошли тесты на гетероскедастичность, нормальность и отсутствие автокорреляции остатков. Недостатком модели можно назвать небольшую положительную автокорреляцию остатков и плохую интерпретируемость одного показателя.

OLS Regression Results

Dep. Variable:	excess_mortality_cumulative_per_million	R-squared:	0.925
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	52.81
Date:	Tue, 23 Aug 2022	Prob (F-statistic):	3.88e-15
Time:	19:06:33	Log-Likelihood:	-297.24
No. Observations:	38	AIC:	610.5
Df Residuals:	30	BIC:	623.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2191.6900	110.221	19.885	0.000	1966.589	2416.791
total_deaths_per_million	0.9480	0.209	4.528	0.000	0.520	1.376
people_fully_vaccinated_per_hundred	-18.5569	8.903	-2.084	0.046	-36.739	-0.375
gdp_per_capita	-0.0288	0.010	-2.956	0.006	-0.049	-0.009
female_smokers	85.0287	18.998	4.476	0.000	46.229	123.829
life_expectancy	-114.6749	45.703	-2.509	0.018	-208.012	-21.338
avg_icu_patients_per_million	-56.4653	12.392	-4.556	0.000	-81.774	-31.157
avg_hosp_patients_per_million	6.7843	2.199	3.086	0.004	2.294	11.275

Omnibus:	0.099	Durbin-Watson:	1.711
Prob(Omnibus):	0.952	Jarque-Bera (JB):	0.225
Skew:	-0.108	Prob(JB):	0.894
Kurtosis:	2.691	Cond. No.	1.60e+04

Рис 6. Модель регрессии.

Процент курящих женщин объяснить достаточно сложно, так как несмотря на то, что курильщики находятся в группе риска, женщины, как правило курят намного меньше мужчин, и соответственно менее подвержены осложнениям. Из выбранных переменных значительную корреляцию – 75% - с процентом курящих женщин имеют только переменные процент людей старше 65 лет и процент людей старше 70 лет. Таким образом, скорее всего существуют более сложные взаимосвязи данных переменных, которые делают процент курящих женщин в стране более значимым в регрессии. Остальные переменные имеют вполне логичную интерпретацию.

Продолжительность жизни, скорее всего, негативно влияет на избыточную смертность по причине того, что коронавирусную инфекцию сложнее переносят люди старшего поколения. Среднее количество пациентов в интенсивной терапии, возможно, имеет отрицательную связь с избыточной смертностью, потому что говорит о большем количестве ресурсов в стране для обеспечения людей лечением. Вакцинация людей, как и в первых моделях значительно влияет на смертность.

Также стоит отметить, что в данном случае нельзя опираться на конкретные значения коэффициентов про интерпретации в связи с тем, многие переменные имеют разный масштаб.

Выводы

В целом вакцинацию можно назвать значительным фактором для снижения как смертности от коронавирусной инфекции, так и для снижения избыточной смертности. Это может быть связано с меньшей нагрузкой на систему здравоохранения. Для снижения смертности внутри страны ограничения властей, связанные с пандемией, не показали однозначного значительного эффекта в обоих исследованиях. Межстрановое исследование выделило 3 дополнительных значимых фактора – продолжительность жизни, ВВП на душу населения и процент курящих женщин. Из них только ВВП на душу населения скорее всего имеет интерпретируемый смысл и связан с большим количеством ресурсов у страны и, возможно, более скорому доступу к вакцинам.

Таким образом, исследование показывает, что вакцинация является самым важным фактором в борьбе со смертность. В отсутствии вакцины значимым фактором для снижения смертности может являться тестирование, но всё же ни один фактор на может сравниться с вакцинацией по значимости.