**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ Федеральное  
государственное автономное образовательное учреждение высшего образования  
Национальный исследовательский университет  
«Высшая школа экономики»**

Факультет экономических наук  
Образовательная программа бакалавривата «Экономика»

**КУРСОВАЯ РАБОТА**  
Построение опережающего индикатора экономической активности   
на основе Google Trends.

Выполнили:  
Макаров Антон Андреевич,   
Альмухаметов Денис Уралович,   
Жуматаев Жанту Зарифуллаевич  
  
Руководитель:  
Мамедли Мариам Октаевна  
PhD, доцент ФЭН  
Департамента теоретической экономики

Москва 2022

**Введение**

В нашем проекте основной целью было построение индикатора оценки ВВП на основе данных Google Trends. Для этого мы использовали данные Google Trends, которые доступны практически на любом интервале времени и которые имеют высокий потенциал для оценки ВВП по этой причине, поскольку появляется возможность оценивать недельные и даже дневные значения ВВП. В данной работе нашей задачей является построение недельного индикатора ВВП.

Работа является актуальной, поскольку в настоящее время ценность информации очень высока, построение достаточно точного прогноза позволит своевременно принимать решения и соответственно иметь конкурентоспособные преимущества на рынке.

**Данные**

Данные, предоставляемые GoogleTrends, принимают значения от 0 до 100 и отражают популярность запроса, темы или какой-либо категории. Шкала соответствует выраженной в процентах доле конкретных запросов в общей массе запросов за определенный период (так, например, периодом может выступать день, неделя, месяц, год и тд). История популярности поисковых запросов составляет около 15 лет, однако этот показатель сильно зависит от конкретной страны.

Отметим несколько важных аспектов относительно категорий и тем. Существует большое количество тем, которые никак не связаны между собой, но при этом поисковый запрос по ним будет идентичным. Например, поисковый запрос 'Apple' можно трактовать и как фрукт, и как название известной технологической компании. В таком запросе уточнение категории поиска позволит избавиться от двусмысленности. В Google Trends также возможно получить данные о популярности конкретной категории среди всех.

Кроме того, важно учитывать язык поискового запроса. Так, например, один и тот же запрос, написанный на разных языках, будет иметь разную популярность, несмотря на одинаковое значение. Это особенно актуально для стран, в которых существует несколько одинаково популярных государственных языков (в качестве примера можно взять Канаду).

Еще одной особенностью данных является повышенный акцент категорий на потребление. Так, значительна по сравнению с другими доля поисковых запросов различных потребительских товаров и услуг. Тем не менее нельзя говорить об отсутствии запросов, отражающих динамику других показателей.

С учётом этих особенностей использование каких-либо конкретных поисковых запросов может оказаться некорректным, поэтому в нашей работе в качестве признаков будут использоваться топики и категории.

Так как эти данные изначально не предназначены для наших целей, использование их в качестве признаков для предсказания ВВП имеет ряд ограничений.

Список использованных категорий можно найти [здесь](https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories), а также в Приложении А.

**Обработка данных**

Результаты нашего исследовании во многом зависят от правильной обработки данных. Собранные сырые данные с сайта [trends.google.com](file:///C:\Users\Zhan\Downloads\Telegram%20Desktop\trends.google.com) представляли из себя временные ряды по категориям и топикам с месячной частотой и охваченным периодом с января 2008 по январь 2022 года. Более ранние года мы не брали в расчет, так как во многих странах мира поисковая система Google тогда еще не имела большой охват и в данных наблюдалась очень высокая дисперсия, которая могла навредить финальной модели.

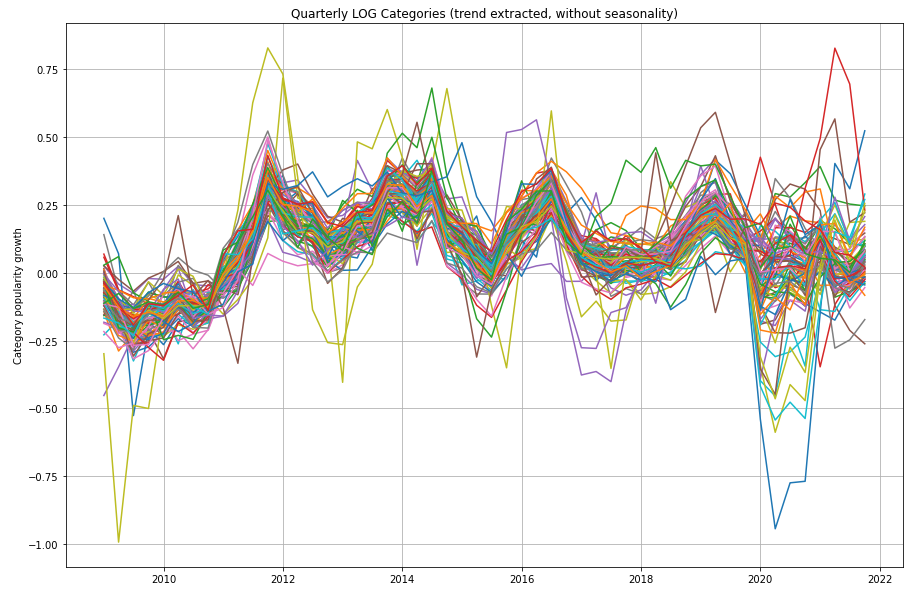
Мы использовали данные по 46 странам, большинство из которых являются странами-участниками OECD, для которых имеются временные ряды квартальной динамики ВВП, которые являются нашей целевой переменной. В выборку не вошли такие страны, как Китай и Саудовская Аравия, так как, во-первых, для них не было достоверных данных по динамике ВВП на сайте OECD за весь рассматриваемый нами период, и, во-вторых, в Китае не так популярна поисковая система Google, поэтому предсказание скорее всего выйдет не самым качественным.

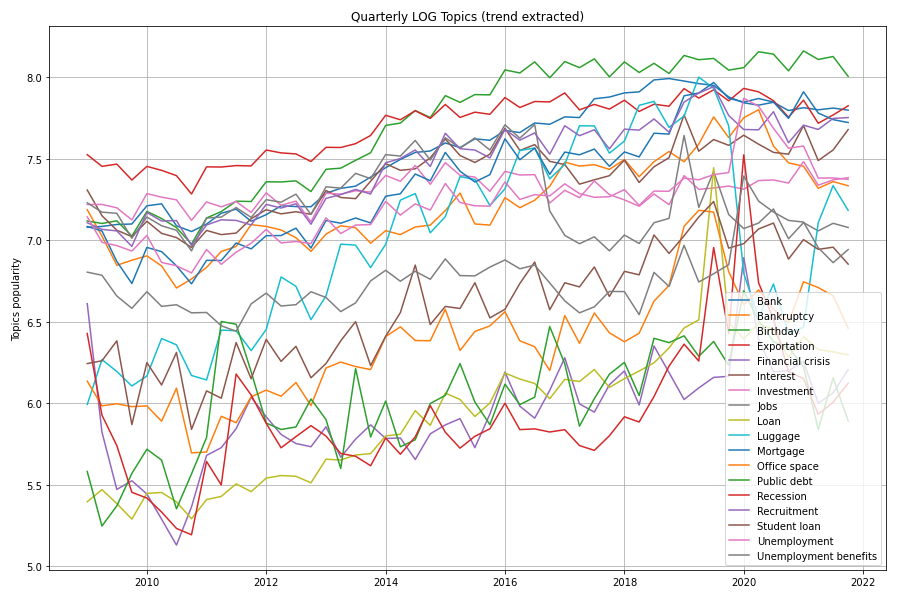
Первоначально в данных необходимо убрать долгосрочный нисходящий тренд. В силу того, что с 2008 года число запросов сильно возросло из-за развития и распространения интернета, а данные представляют из себя популярность запроса, которая считается как доля в общем числе запросов за период, то все ряды имеют нисходящий тренд, который необходимо извлечь. Для этого на логарифмированных рядах высчитывается первая главная компонента (PCA) и масштабируется под среднее и стандартное отклонение логарифмированного ряда. Предполагается, что PC1 способна уловить эффект нисходящего тренда из-за увеличения знаменателя в индикаторе популярности запроса.

Далее необходимо учесть сезонность, которая более свойственна категориям, нежели чем для топиков. Для последних мы берем просто логарифм, тогда как для категорий мы берём прирост популярности, взяв для каждого месяца разность между логарифмированным индикатором в текущем месяце и тем же месяцем в прошлом году. В итоге, а обучающую выборку попадут данные начиная с 2009 года.

В 2011 и в 2016 годах в рядах наблюдаются разрывы, которые связаны с изменением методики подсчета индикатора google trends. Мы вычитаем прирост в этих годах из последующих периодов, чтобы выровнять данные и сделать их более непрерывными, таким образом изменения не имеющие отношения к популярности запросов не повлияют на нашу модель.

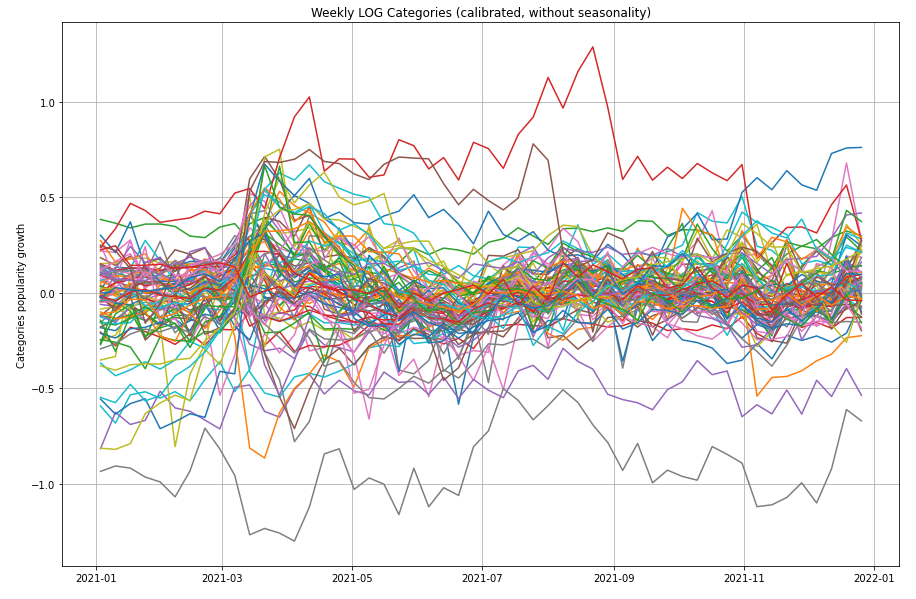
В конце мы усреднили месячные данные по кварталам, так как целевая переменная прироста ВВП имеется только в квартальной частоте.

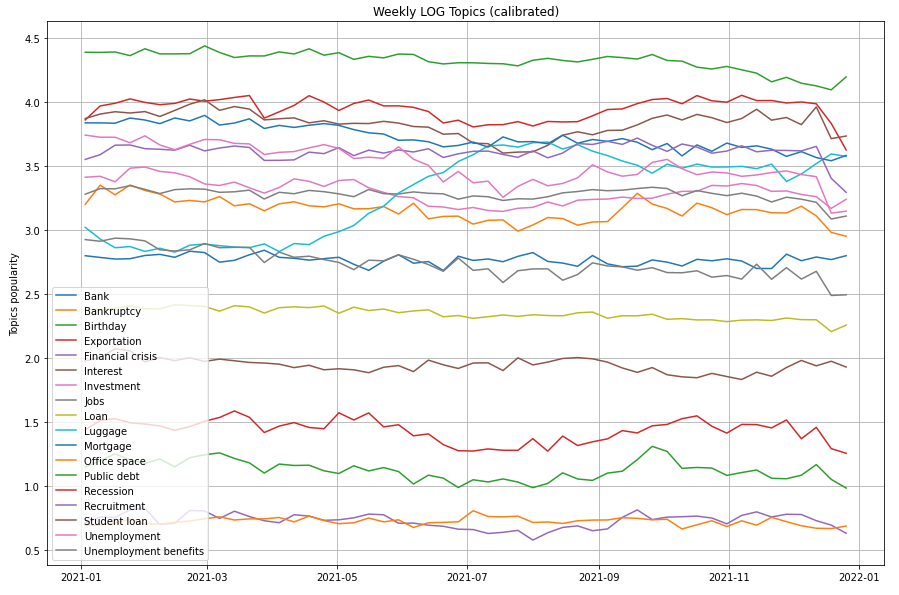




Для построения недельного индикатора мы используем недельные данные популярности запросов по топикам и категориям за период с 2020 по 2021 год. Самая главная задача в обработке недельных данных заключалась в учете различного масштаба месячных и недельных данных. Требовалось калибровка недельных данных под месячные, которую мы осуществили в два шага. Сначала мы интерполировали месячные данные на недельные частоты и затем умножили месячные ряды на калибровочную константу, которая была уникальная для каждой страны и категории/топика. Константа высчитывалась как среднее отношение между интерполированными месячными данными и изначальными недельными данными. Таким образом, мы нивелировали эффект разного масштаба, что обеспечит более стабильное предсказание нашей модели.

На недельных данных не наблюдается долгосрочный тренд из-за рассматриваемого периода в один год, поэтому извлечение тренда здесь не требуется. Тогда как сезонность убирается тем же образом, что и для месячных данных, взятием лог-разности с 2020 годом.





В конечном счете мы постарались учесть все возможные шоки, не связанные с реальной динамикой популярности запросов, чтобы модель показала наилучшее качество предсказания.

**Моделирование**

В качестве основной модели используется полносвязная нейронная сеть с двумя слоями. Так как данные Google Trends скорее имеют dense природу[[1]](#footnote-1), было необходимо включить множество разных категорий, запросов и топиков. В связи с этим, предпочтение было отдано общей модели на данных всех стран, а не множеству моделей для отдельных стран.

В данные были добавлены дамми переменные для каждой страны. Предполагалось, что модель сможет уловить зависимость между данными Google Trends и приростом ВВП для каждой страны благодаря своей нелинейной природе.

Так как полносвязные сети сильно зависят от инициализации, было решено обучить модель 6 раз при разных random seed и взять среднее предсказаний для относительной устойчивости результатов, как это было сделано в (Nicolas Woloszko. 2020).

Оптимизация гиперпараметров бы произведена на валидационной выборке. Она заключалась в стандартном переборе числа нейронов в первом и втором слоях, а также величины learning rate. Выбор лучшей комбинации гиперпараметров осуществлялся на основе следующей метрики: - H(mean\_20, mse\_tot), где H – гармоническое среднее, mean\_20 – средняя MSE 20 стран с наименьшей MSE на валидации, mse\_tot – MSE на всех странах. Такая метрика была придумана для того, чтобы сконцентрироваться на странах с уже неплохими результатами вместо того, чтобы улучшать качество на странах с плохим перформансом модели. Такой подход помогает избежать случаев, в которых мы будем отдавать предпочтение наборам гиперпараметров, при которых ошибка на странах с наибольшей MSE будет ниже, чем обычно, но качество на лучших странах будет заметно хуже, при этом общая MSE уменьшится.

Хоть в статье (Nicolas Woloszko. 2020) и была приведена исчерпывающая аргументация в пользу полносвязной сети, мы хотели убедиться, что из довольно популярных (в случае временных рядов) моделей не найдется лучшей.

В качестве первой из таких моделей мы выбрали градиентный бустинг. Эта модель является очень мощной, так как будучи основанной на решающих деревьях она способна находить почти любые нелинейные зависимости в данных. Однако у градиентного бустинга над деревьями есть существенный недостаток: он не может экстраполировать данные, что существенно при его использовании на временных рядах. Так как предсказания решающих деревьев – это значения, попавшие в листы без каких-либо преобразований над ними, максимум усредненные, то итоговая композиция над такими деревьями будет хороша только если в тренировочной выборке уже был похожий паттерн. Другими словами, градиентный бустинг над деревьями хорош в интерполяции.

Но в силу того, что во временных рядах часто есть тренд, то градиентный бустинг не самая подходящая модель для предсказания на временных данных. Решить эту проблему можно, если добавить в обучающую выборку признаки, связанные с временными рядами, например, лаги, частоты, периоды.

Еще одним объяснением может послужить тот факт, что бустинг для композиции использует неглубокие деревья, но как уже утверждалось наши данные имеют dense природу, то есть целевая переменная зависит сразу от многих признаков, а построение дерева по случайному их подмножеству и дальнейшее усреднение не может дать хороший результат.

Также были рассмотрены несколько статистических моделей, которые предназначены для работы с рядами. Особенностью применения статистических моделей являлся тот факт, что они не предназначены для работы с повторяющимися временными интервалами, поэтому в отличие от бустинга и нейронной сети статистическая модель обучалась конкретно для каждой страны, а не для всех данных.

Тем не менее, как показали тесты, наилучший результат все равно показывает нейронная сеть, в чем мы можем убедиться в результатах.

**Результаты**

1. Статистические модели

Для удобства, компактности и иллюстративности мы посмотрим на результаты 10 лучших странах и отдельно на РФ для каждой модели.

Среди статистических моделей была рассмотрена модель DLT (Damped Local Trend), обученная на всех данных по одной стране или только на 10 самых скоррелированных с целевой переменной столбцов, учитывающая и не учитывающая сезонность соответственно. Ниже представлены ее результаты.

**10 лучших стран после обучения на всех столбцах с учетом сезонности:**

|  |  |
| --- | --- |
| Страна | MSE |
| Болгария | 2.29 |
| Финляндия | 7.82 |
| США | 10.93 |
| Индонезия | 16.83 |
| Нидерланды | 17.91 |
| Словакия | 20.91 |
| Латвия | 21.99 |
| Литва | 22.26 |
| Швеция | 34.80 |
| Исландия | 40.13 |

**10 лучших стран после обучения на 10 самых коррелированных столбцах с учетом сезонности:**

|  |  |
| --- | --- |
| Страна | MSE |
| Южная Корея | 4.24 |
| Финляндия | 5.15 |
| Дания | 6.58 |
| Швейцария | 9.52 |
| Швеция | 9.68 |
| Латвия | 11.43 |
| Германия | 13.32 |
| США | 15.50 |
| Люксембург | 17.37 |
| Словакия | 21.03 |

**10 лучших стран после обучения на 10 самых коррелированных столбцах без учета сезонности**

|  |  |
| --- | --- |
| Страна | MSE |
| Южная Корея | 5.28 |
| Финляндия | 7.04 |
| Латвия | 7.96 |
| Дания | 9.09 |
| Швеция | 10.65 |
| Швейцария | 11.64 |
| США | 14.48 |
| Нидерланды | 14.62 |
| Германия | 15.06 |
| Израиль | 16.53 |

Кроме этой модели была также рассмотрена базовая модель Exponential Smoothing, которая также работает с трендом. Правда она является более простой моделью по сравнению с DLT, поэтому ее результаты оказались хуже как для 10 лучших стран, так и среднее по всем странам в целом.

|  |  |
| --- | --- |
| Страна | MSE |
| Южная Корея | 22.23 |
| Норвегия | 24.09 |
| Латвия | 28.15 |
| Финляндия | 32.90 |
| Ирландия | 34.37 |
| Швейцария | 35.45 |
| Россия | 37.94 |
| Словакия | 38.95 |
| Дания | 41.83 |
| Швеция | 43.94 |

Последней статистической моделью, которая была использована, является ARIMA, которая, однако, не принимает на вход регрессоры, по этой причине результаты модели значительно хуже, чем у DLT.

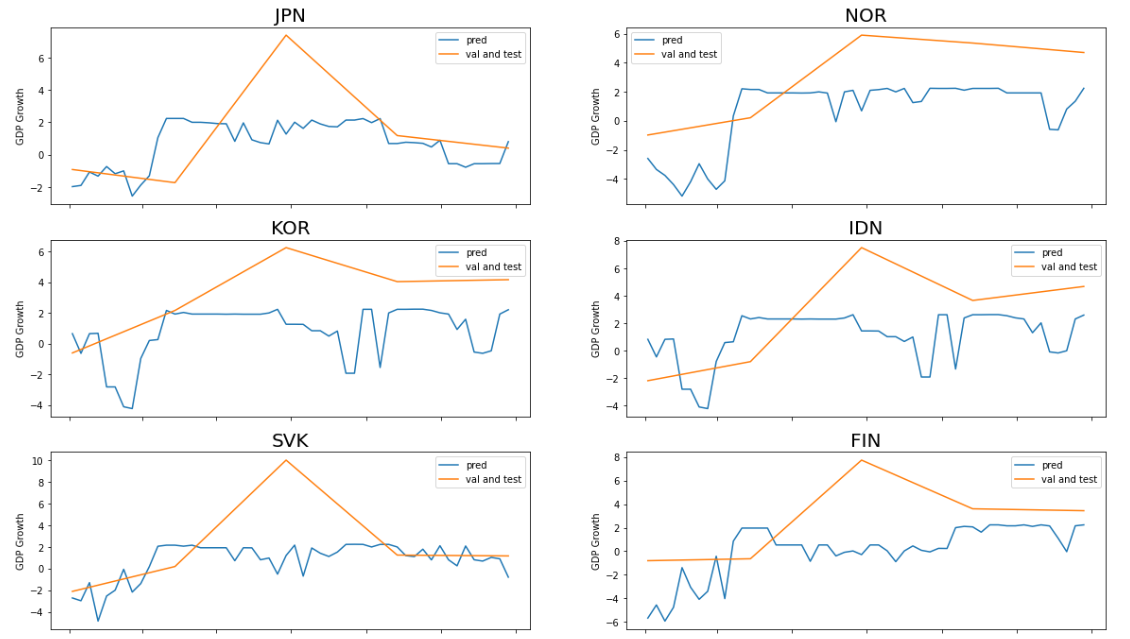
|  |  |
| --- | --- |
| Страна | MSE |
| Финляндия | 4.91 |
| Россия | 9.60 |
| Швейцария | 10.00 |
| Словакия | 12.06 |
| Латвия | 12.87 |
| Эстония | 12.91 |
| Дания | 14.79 |
| Южная Корея | 15.25 |
| Норвегия | 19.79 |
| Латвия | 20.29 |

Таким образом, наилучшей из статических моделей стала модель DLT, причем разница между этой моделью и другими достаточно существенна. Тем не менее, даже она сильно уступает по качеству нейронной сети.

1. Gradient Boosting

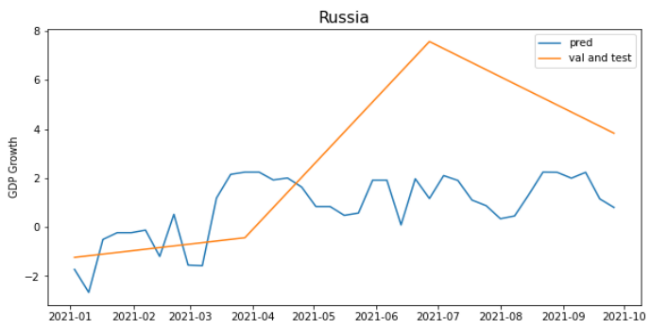
Таблица с MSE на 10 лучших стран и России по предсказанию:

|  |  |
| --- | --- |
| Страна | MSE |
| Япония | 5.83 |
| Норвегия | 10.65 |
| Корея | 11.18 |
| Россия | 11.24 |
| Индонезия | 11.48 |
| Словакия | 13.15 |
| Финляндия | 15.62 |
| Швейцария | 15.84 |
| Чехия | 16.75 |
| Болгария | 18.41 |
| Литва | 19.23 |

Графическое представление предсказанного недельного индикатора против реальных данных по квартальному приросту ВВП за 2021 год: 



Графическое представление для России:

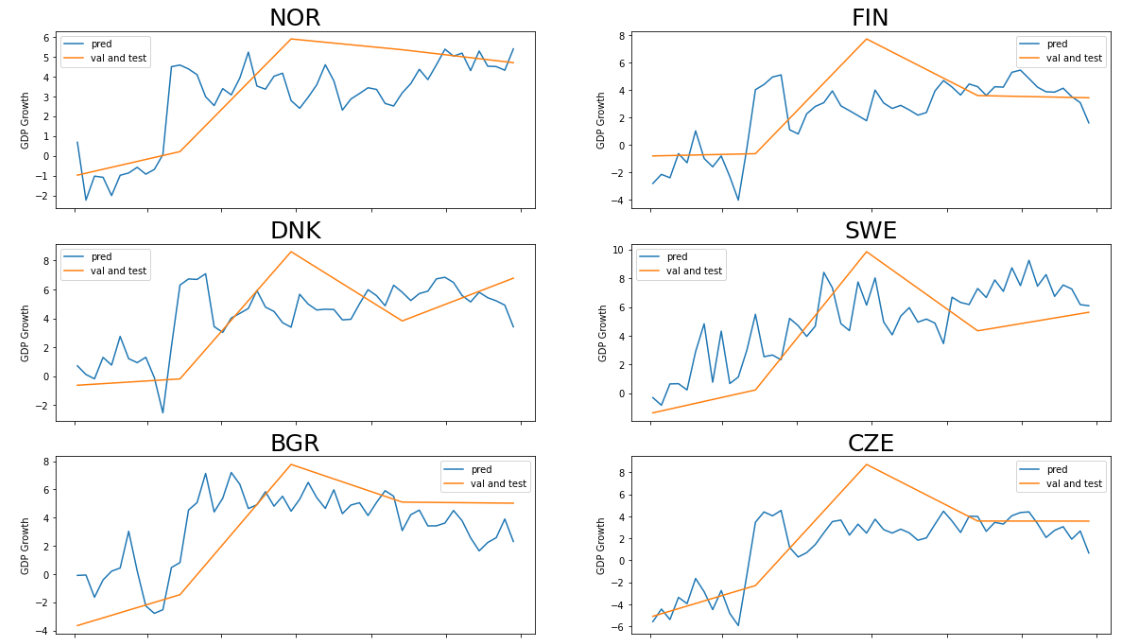


Как мы можем убедиться, бустинг, хоть и значительно лучше статистических моделей, но тем не менее не до конца справляется с задачей. На графиках хорошо видна проблема градиентного бустинга с экстраполяцией. Предсказанный ряд «упирается в потолок», так как в тренировочной выборке не было таких квартальных приростов. К тому же 2021 год характеризуется восстановительным ростом во всех странах после пандемии, поэтому модель, обученная на небольших приростах ВВП в прошлом, что особенно характерно для развитых стран, плохо себя показывает в 2021 году.

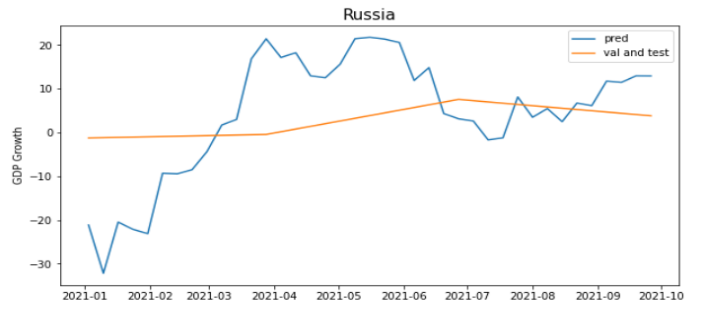
1. MLP Regressor (Fully connected network)

Таблица с MSE на 10 лучших стран и России по предсказанию:

|  |  |
| --- | --- |
| Страна | MSE |
| Норвегия | 3.5 |
| Финляндия | 5.9 |
| Дания | 6.07 |
| Швеция | 6.45 |
| Болгария | 6.6 |
| Чехия | 7.93 |
| Латвия | 8.51 |
| Австрия | 8.86 |
| Польша | 9.75 |
| Германия | 11.44 |
| Россия | 13.72 |

Графическое представление предсказанного недельного индикатора против реальных данных по квартальному приросту ВВП за 2021 год: 

Графическое представление для России:



Как мы можем убедиться, смотря на показатели качества и графики, нейронная сеть показала себя лучше, чем остальные модели. В данном случае уже не наблюдается той проблемы, которая была у градиентного бустинга – это видно на графиках. Как и предполагалось, нейронная сеть смогла уловить отличия между странами, благодаря своей нелинейной природе. Это можно проследить, например, на графиках для Норвегии и Финляндии, где четко видно, как построенный индикатор пытается аппроксимировать реальную зависимость.

Среди причин достаточно больших ошибок (MSE) можно выделить следующие:

1. Недостаточное количество признаков
2. Некоторые из выбранных признаков являются излишне зашумленными или смещенными. Такое может происходить для всех признаков в рамках одной страны, например, в случае слабой популярности поисковика Google в данной стране.
3. Ошибки посчитаны с учетом “ненастоящих” данных. Понедельного прироста ВВП мы не знаем, а располагаем только интерполяцией, что существенно может повысить ошибку.

С уверенностью говорить о качестве построенного индикатора невозможно. Графики предсказаний и настоящего прироста показывают лишь, что модель повторяем во всяком случае направление, что позволяется с какой-то долей уверенности говорить о том, что индикатор построен относительно успешно. Сравнивая графики и MSE с другими моделями, можно прийти к этому же выводу.

**Вывод**

В результате нашей работы мы собрали и обработали сырые данные Google Trends и построили недельный индикатор экономической активности с помощью нескольких моделей. У каждой модели были свои плюсы и минусы, но в итоге, сравнивая результаты всех моделей, можно сказать, что полносвязная нейронная сеть наилучшим образом справляется с поставленной задачей.

Предсказания для многих развитых стран вышли довольно хорошими, особенно на фоне кризисных явлений 2020 и 2021 годов, повторяя динамику поквартального прироста ВВП. Однако полноценно судить о качестве модели нельзя в силу того, что мы использовали интерполяцию поквартальных данных на недельные частоты.

Приложение А. Список использованных топиков и категорий.

Topics:

['Bank', 'Bankruptcy', 'Birthday', 'Exportation',

'Financial crisis', 'Interest', 'Investment', 'Jobs', 'Loan',

'Luggage', 'Mortgage', 'Office space', 'Public debt', 'Recession',

'Recruitment', 'Student loan', 'Unemployment',

'Unemployment benefits']

Categories:

['Advertising & Marketing', 'Agricultural Equipment',

'Agriculture & Forestry', 'Alcoholic Beverages', 'Apparel',

'Aquaculture', 'Architecture', 'Auto Financing',

'Autos & Vehicles', 'Aviation', 'Bankruptcy', 'Boats & Watercraft',

'Building Materials & Supplies', 'Business Services',

'Chemicals Industry', 'Civil Engineering', 'College Financing',

'Commercial Lending', 'Computer Hardware',

'Computers & Electronics', 'Construction & Maintenance',

'Construction Consulting & Contracting', 'Consumer Electronics',

'Credit & Lending', 'Developer Jobs', 'Distribution & Logistics',

'Education', 'Emergency Services', 'Entertainment Industry',

'Entertainment Media', 'Events & Listings', 'Finance',

'Fire & Security Services', 'Food & Drink', 'Food Production',

'Footwear', 'Forestry', 'Freight & Trucking',

'Fuel Economy & Gas Prices', 'Gifts & Special Event Items',

'Government', 'Grocery & Food Retailers', 'Health Insurance',

'Home & Garden', 'Home Appliances', 'Home Financing',

'Home Furnishings', 'Home Improvement',

'Hospitals & Treatment Centers', 'Hotels & Accommodations',

'Import & Export', 'Industrial Materials & Equipment', 'Insurance',

'Internet & Telecom', 'Jobs', 'Legal Services', 'Luxury Goods',

'Mail & Package Delivery', 'Maritime Transport',

'Medical Facilities & Services', 'Mental Health',

'Office Supplies', 'Performing Arts', 'Pharmaceuticals & Biotech',

'Pharmacy', 'Printing & Publishing',

'Professional & Trade Associations', 'Programming',

'Rail Transport', 'Real Estate', 'Real Estate Agencies',

'Restaurants', 'Retail Trade', 'Shopping', 'Social Services',

'Sports', 'Tobacco Products', 'Transportation & Logistics',

'Travel', 'Travel Agencies & Services',

'Unwanted Body & Facial Hair Removal', 'Vehicle Brands',

'Vehicle Licensing & Registration', 'Veterinarians']

1. (Giannone, Lenza and Primiceri, 2018) [↑](#footnote-ref-1)