ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И
ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ
ФЕДЕРАЦИИ»
(РАНХиГС)

А.В. Зубарев, Е.А. Голованова

ИСПОЛЬЗОВАНИЕ GOOGLE TRENDS ДЛЯ ПРОГНОЗИРОВАНИЯ: ОБЗОР И ПРИМЕНЕНИЕ ДЛЯ ПРОГНОЗИРОВАНИЯ РОЗНИЧНЫХ ПРОДАЖ

Москва 2021

Аннотация

В связи с ростом популярности Интернета множество покупок делается в онлайн магазинах. Сервис Google Trends собирает данные по запросам пользователей и формирует из них категории. В данной работе мы обозреваем существующие методы для прогнозирования с помощью данного сервиса, а также предпринимаем попытку спрогнозировать динамику розничных продаж, используя макроэкономические переменные и категории в Google Trends, соответствующие различным товарным группам продовольственных и непродовольственных товаров. Для каждого вида ретейла мы строим лучшие прогнозные модели из макроэкономических переменных и пытаемся их улучшить путем добавления трендов.

Abstract

Due to the growing popularity of the Internet, many purchases are made in online stores. The Google Trends service collects data based on user requests and breaks them down into categories. In this paper, we review the existing forecasting methods using this service, and make an attempt to predict the dynamics of retail sales using macroeconomic variables and categories in Google Trends corresponding to various commodity groups of food and non-food products. For each type of retail, we build the best predictive models from macroeconomic variables and try to improve them by adding trends.

А.В. Зубарев к.э.н., старший научный сотрудник Лаборатории математического моделирования экономических процессов ИПЭИ РАНХиГС

Е.А. Голованова младший научный сотрудник Студенческого центра экономических исследований ИПЭИ РАНХиГС

Данная работа подготовлена на основе материалов научно-исследовательской работы, выполненной в соответствии с Государственным заданием РАНХиГС при Президенте Российской Федерации на 2021 год

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Глобальная пандемия, приведшая к значительному снижению спроса в мировой экономике и вынужденному карантину, напрямую сказалась на реальных денежных доходах в России, а во втором квартале 2020 привела к их рекордному с начала века снижению [1]. Такая динамика явным образом влияет на расходы населения, важную долю в которых занимают траты на продовольственные и непродовольственные розничные товары. В связи с этим интересно выявить переменные, потенциально объясняющие динамику розничных продаж и помогающих их прогнозированию. Помимо стандартных макроэкономических показателей важным инструментом могут являться поисковые запросы по отдельным товарным группам, которые помогают отследить динамику интереса пользователей. Хочется понять, насколько данные о поисковых запросах помогают прогнозировать розничные продажи. Для решения данной задачи мы рассматриваем регрессии с некоторым базовым набором ключевых макроэкономических показателей, объясняющих продовольственный и непродовольственный ретейл, а затем используем Google trends для улучшения прогнозный свойств различных моделей.

Структура работы выглядит следующим образом. Во втором разделе мы приводим обзор исследований из области выявления факторов и прогнозирования потребления, в которых использовались Google Trends. В третьем разделе приведены данные, которые используются для расчетов. В четвертом и пятом разделах проверяется влияние поисковых запросов в Google на качество среднесрочных и краткосрочных прогнозов розничных продаж. В последнем разделе мы приводим выводы из полученных результатов.

1 Обзор литературы

Использование Google Trends в исследованиях – сравнительно новая тенденция. Однако на сегодняшний день использование Google Trends в различных целях часто встречается в научной литературе. Это объясняется относительной простотой получения и обработки этих данных, их удобным и понятным представлением. Кроме того, поскольку Интернетом и браузером Google пользуется огромное число людей, такие запросы в совокупности составляют репрезентативные индикаторы интереса пользователей. Таким образом, сервис Google Trends содержит огромную базу различных индикаторов, в том числе способных описывать поведение потребителей.

Среди первых исследований с использованием трендов в области потребления выделяется работа [2], где предпринимаются попытки наукастинга розничных продаж, продаж автомобилей, домов и путешествий. Авторы используют категории из Google Insights for Search, впоследствии объединенного с Google Trends. Авторы подчеркивали, что предсказание настоящего полезно, поскольку оно может помочь определить «поворотные

моменты» в экономических временных рядах. Например, если люди начнут выполнять значительно больше запросов по запросу «Агенты по недвижимости» в определенном месте, можно будет делать предположения, что в ближайшем будущем продажи домов в этом районе могут вырасти. Vossen and Schmidt [3] впервые попробовали использовать данные Google Trends в качестве показателя для прогнозирования частного потребления в США. Авторы заключили, что в сравнении с индексами, основанными на опросах, такими как индекс потребительских настроений Мичиганского университета[1] и индекс потребительского доверия Conference Board[2], главные компоненты, выделенные из категорий Google Trends, помогают получить более точные прогнозы.

На основе идей из выше представленных работ было создано множество исследований, подтверждающих полезность Google Trends для улучшения прогнозов потребления. По мере роста активных пользователей Интернета все больше покупателей используют его для сбора информации о товарах и выбора покупок. Таким образом, в Google Trends становится больше информации, важной для прогнозирования. Среди основных направлений в этой области можно выделить прогнозирование розничных продаж, крупных покупок, услуг и совокупного потребления.

Говоря о прогнозировании в области продуктового ретейла, следует упомянуть работу [4] , где Google Trends используются для прогнозирования спроса на закуски. Авторы разбили все закуски на подгруппы и подобрали к ним ключевые слова в трендах. Каждая подгруппа трендов по ключевым слова помогала прогнозировать розничные продажи с улучшениями MAPE на 2.2–7.66%. Также тренды используются для анализа поведения потребителей продуктовых магазинов в коммерческих целях Например, Кухлер et al. [5] провели исследование на тему знают ли покупатели разницу между маркировками «натуральной» и «органической» еды и оценили, влияет ли это на продажи.[5]). Для предсказания fashion-ретейла, например, в статье [6] в модель для нейронной сети добавлен тренд по слову «Burberry» для предсказания продаж одноименного магазина люксовой одежды. В работе [7] использовался тренд «Winter jacket in NYC» и применялись дисперсионный анализ и корреляция Пирсона для обнаружения закономерности поведения потребителей при поиске сезонной одежды в зимние месяцы. В работе [8] автор использует данные по агрегированному ретейлу в Норвегии и прогнозирует его с помощью 51 категории и 148 ключевых слов в Google Trends.

В исследованиях, связанных с крупными покупками, тренды используются для прогнозирования цен на жилье и автомобили. Wu и Brynjolfsson [9] прогнозировали тенденции на рынке недвижимости посредством включения трендов наряду с эндогенными

[1] https://fred.stlouisfed.org/series/UMCSENT
[2] https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

переменными, такими как индекс цен на жилье, объем продаж жилья. В работе [10] использовались комбинации из категорий, ключевых слов в области рынка жилья и смежного с ним рынка строительства. Wijnhoven и Plant [11] обнаружили, что Google Trends более эффективны для прогнозирования на рынке автомобилей, чем информация из постов в социальных сетях. Кроме того, не было замечено разницы в корреляции цен дорогих и дешевых автомобилей с Google Trends.

Данные запросов поисковых систем хорошо отражают интерес в области туризма и полезны для прогнозирования потребления туристических продуктов. Li et al. [12] прогнозировали спрос в отрасли туризма в Пекине, используя Google Trends. Авторы предложили процедуру создания составного поискового индекса, используя обобщенную динамическую факторную модель (GDFM). Выбор модели объясняется тем, что соответствующие составляющие отрасли туризма, такие как спрос на услуги ресторанов, гостиниц и туристических агентств, тесно коррелируют друг с другом. Опираясь на стандартные запросы перед путешествиями, как еда, погода, отели, авторы разбили их на категории и подобрали к ним ключевые слова. В качестве бенчмарков использовались модели ARMA(1,1) с индексом, созданным методом главных компонент и без него. Полученные результаты свидетельствуют о том, что предложенный метод повышает точность прогноза месячных объемов туризма в Пекине по сравнению с двумя бенчмарками. Похожую работу проделали Padhi и Pati [13] для прогнозирования туризма в Индии. В статье Dergiades, Mavragani, Pan [14] учитывается то, что запросы по туристическим продуктам в одну страну поступают на разных языках и с разных поисковых платформ. С помощью двух процедур тестирования на отсутствие причинно-следственных связей выявлено, что с учетом этих двух факторов индикатор интенсивности поиска лучше прогнозирует количество иностранных посетителей. Помимо туризма Google Trends используется для предсказания спроса на косметологические услуги. Анализ, приведенный в статье [15], подтверждает жизнеспособность этого инструмента для оценки интереса пациентов к нехирургическим косметическим процедурам.

Вышеперечисленные исследования концентрируются на решении конкретных коммерческих задач, будь то продажи домов или поход к косметологу. Прогнозирование совокупного потребления является более интересной с макроэкономической точки зрения задачей, однако и здесь уместно использование динамики интереса потребителей. Fasulo et al. [16] прогнозировали расходы домашних хозяйств на потребление в Италии с использованием Google Trends, связанных с ключевыми словами на тему расходов. Scott и Varian [17] использовали категории Google Trends для наукастинга индекса потребительских настроений Мичиганского университета. Из 151 категории отбор предикторов производился с помощью байесовского подхода. В работе [18] авторы изучают вклад данных Google Trends

в предсказание частного потребления в США. Для этого используются данные по потреблению товаров долгосрочного и краткосрочного пользования, а также услуг. Для каждой категории товаров строится линейная регрессия, включающая макроэкономические переменные, такие как индексы потребительского доверия, индекс волатильности, реальный располагаемый доход и ставка по трехмесячной гособлигации. Авторы используют категоризации из Bureau of Economic Analysis, которая включает в себя слова, связанные с каждой категорией потребления товаров [3]. Данные ключевые слова сопоставляются с категориями в Google Trends. Из полученных трендов выделяются главные компоненты. Также авторы помимо категорий рассматривают запросы по ключевым словам «рецессия» и «увольнение». Затем предпринимается попытка спрогнозировать потребление, добавляя в модель либо компоненты, либо запросы по ключевым словам, либо и то, и другое. Авторы подтверждают, что тренды помогают лучше прогнозировать изменения в потреблении. В работе [19] проверяется, содержат ли социальные сети полезные сигналы о будущих потребительских расходах помимо тех, которые содержатся в макроэкономических переменных, обычно используемых для их прогнозирования. В ходе исследования было использовано несколько методов получения количественных предикторов из выражений покупательских намерений, встречающихся в тексте публичных постов в социальных сетях (Twitter, Facebook). Методы основаны на обнаружении слов, относящихся к предполагаемым покупкам, а затем построении их семантических представлений, эмбеддингах (преобразования в числовой вектор) и кластеризации по их значению. Семантические предикторы оценивались путем включения их наряду с макроэкономическими переменными в модели для индекса потребительских расходов[3] (Consumer Spending Index). Также в исследовании использовались предикторы, построенные на основе данных Google Trends. Было подтверждено, что и посты в социальных сетях, и данные Google Trends по отдельности помогают значительно улучшить прогнозы потребительских расходов по сравнению с моделями, где используются только макроэкономические переменные. Автор также попробовал совместить в одной модели тренды и посты в социальных сетях, но результаты получились неоднозначными: в некоторых экспериментах эта комбинация переменных приводила к значительному снижению ошибок, в то время как в других не было замечено улучшения.

На основании проведённого анализа литературы можно заключить, что использование данных о поисковых запросах в интернете в виде Google Trends позволяет существенно улучшить прогнозную силу моделей для множества разнообразных экономических и не только показателей. В нашем исследовании в качестве примера

[3] http://www.gallup.com/poll/112723/gallup-daily-usconsumer-spending.aspx

использования такой методики мы предпринимаем попытку оценить значимость различных Google Trends для прогнозирования объёма розничных продаж. Используя агрегированные данные по ретейлу, мы можем отдельно изучить спрос на продовольственные и непродовольственные товары в российской экономике. Для проведения нашего анализа мы используем категории в Google Trends и рассматриваем важные макроэкономические переменные, также потенциально характеризующие динамику розничных продаж.

2 Данные

Мы рассматриваем промежуток с января 2015 по февраль 2021 года включительно. В текущем исследовании использованы ежемесячные данные Росстата по выручке с совокупных розничных продаж. Мы используем индексы физического объема оборота продовольственной и непродовольственной розничной торговли как целевые переменные. Ряды Google Trends логарифмируются и рассматриваются в виде месяц к аналогичному месяцу предыдущего года (сезонная разность) как в работах [3], [18], что помогает избежать корректировки сложного характера сезонности в этих рядах.

Также рассмотрение данных в таком виде позволяет возможную проблему нестационарности, вызванную насыщением Интернета пользователями на рассматриваемом промежутке. В приложении А приведены результаты расширенного теста Дики-Фуллера на наличие единичных корней в рассматриваемых рядах. Для продовольственных товаров нулевая гипотеза о наличии нестационарности не отвергается для 4 из 7 категорий трендов, а для непродовольственных – для 21 из 25.

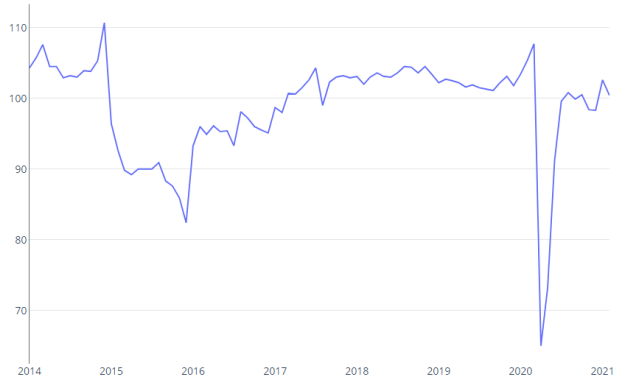
В качестве макроэкономических переменных по аналогии с исследованием [18] мы рассматриваем российский индекс волатильности для контроля за изменениями потребления в связи с колебаниями на фондовом рынке. Также мы включаем в модели индикатор потребительского доверия от ВЦИОМ, который показывает, насколько благоприятно по мнению россиян текущее время для совершения крупных покупок. Чем выше значение индекса, тем более благоприятным россияне считают текущий момент для крупных приобретений[4].

Мы берем в рассмотрение цену на нефть марки Brent, так как для России цена на нефть оказывает существенное влияние на экономику, даже если изменения этой цены вызваны шоками глобального спроса (см. [20], [21]), а в качестве альтернативного показателя мы рассматриваем реальный эффективный девизный валютный курс, динамика которого в существенной степени связана с ценой на нефть. Данные по цене на нефть были скорректированы на американский индекс потребительских цен, очищенный от сезонности.

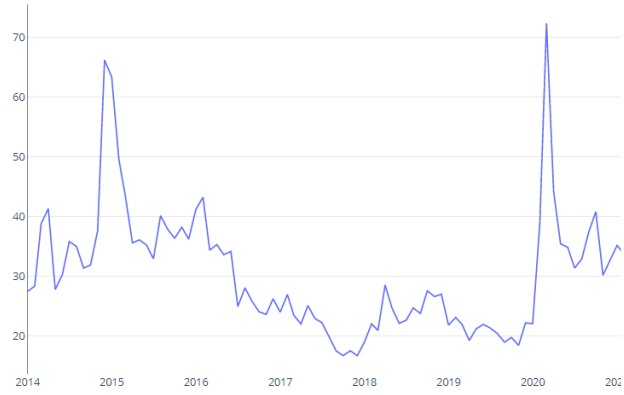
[4] Примечательно, что с 2015 по 2017 годы в данных наблюдаются пропуски, которые мы заполнили линейной интерполяцией.

В приложении Б представлен перечень переменных со ссылками на источники. Мы приводим все данные месяц к месяцу предыдущего года и рассматриваем ряды в логарифмах. На рисунке 1 изображена динамика перечисленных макроэкономических переменных. Примечательно, что в апреле 2020 года из-за последствий пандемии и повсеместной практики введения карантина можно наблюдать минимальные значения у непродовольственного ретейла, у цены на нефть и у индекса потребительского доверия.

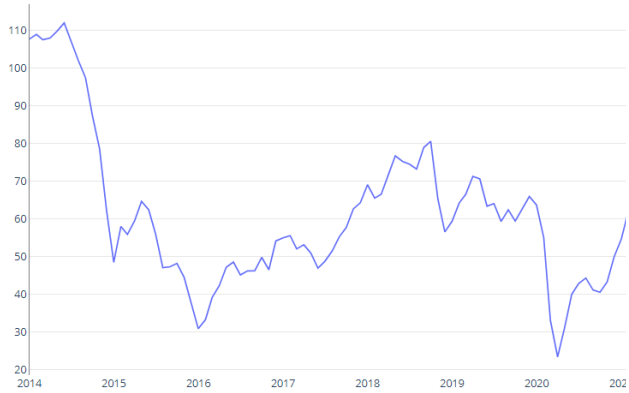
А. Продовольственный ретейл, %



Б. Непродовольственный ретейл, %

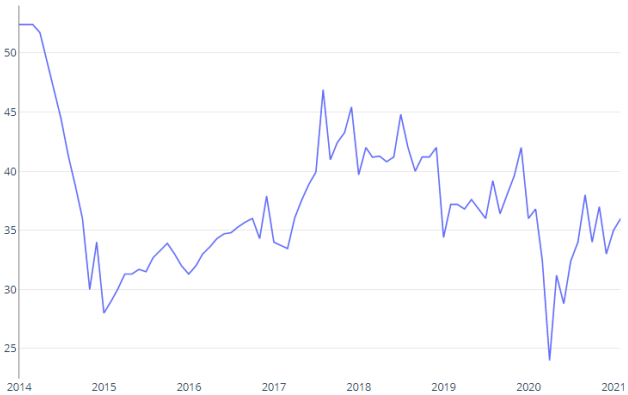


В. Индекс волатильности RVI, %



Г. Цена на нефть (Brent), долл. США

Д. Реальный валютный курс, долл. США/руб.



Е. Индекс потребительского доверия, пункты

Рисунок 1 – Динамика макроэкономических переменных

В качестве трендов для продовольственного ретейла мы взяли в рассмотрение Alcoholic Beverages, Candy & Sweets, Fruits & Vegetables, Meat & Seafood, Baked goods, Coffee

& Tea, Tobacco products. Для непродовольственного ретейла мы выбрали следующие тренды: Textiles & Nonwovens, Children's clothing, Men's clothing, Women's clothing, Undergarments, Footwear, Cleaning Supplies & Services, Make-Up & Cosmetics, Perfumes & Fragrances, Computer Hardware, Mobile phones, Audio equipment, Rugs & Carpets, Sofas & Chairs, Home Storage & Shelving, Construction & Power Tools, Disabled & Special Needs, Drugs & Medications, Book Retailers, Magazines, Bicycles & Accessories, Motorcycles, Classic Vehicles, Vehicle Fuels & Lubricants, Major Kitchen Appliances. Мы решили использовать категории, так как это поможет выя вить общий интерес к товарам на основе разнообразных запросов. В случае ключевых слов не совсем очевидно, какие запросы нужно использовать, и велика вероятность не учесть какие-то запросы, которые помогли бы составить общую картину интереса к этим товарам. Таким образом, для продуктового ретейла у нас подобрано 7 трендов, а для непродовольственного – 25[5]

Google Trends предоставляет относительные частоты для выбранной категории и в зависимости от выбранного периода масштабирует данные от 0 до 100. Максимальное значение соответствует наибольшему числу запросов, в то время как все остальные значения масштабируются к этому максимуму. Все выбранные тренды мы скачиваем с 2004 года для получения истинной динамики ряда. Уточняем, что данные берутся для России и по поисковым запросам в Интернете (Web Search). Затем мы обрезаем ряд с 2015 года для сопоставимости вычислений. Все перечисленные выше макроэкономические переменные, за исключением индекса потребительского доверия, мы также логарифмируем и берем в терминах месяц к аналогичному месяцу предыдущего года. На рисунках 2 и 3 представлены продовольственные и непродовольственные категории Google Trends.

[5] Мы предпринимали попытку рассмотрения моделей с главными компонентами из этих категорий товаров, однако ввиду чувствительности исторической динамики главных компонент к добавлению новых данных мы отказались от идеи их использования для прогнозирования.

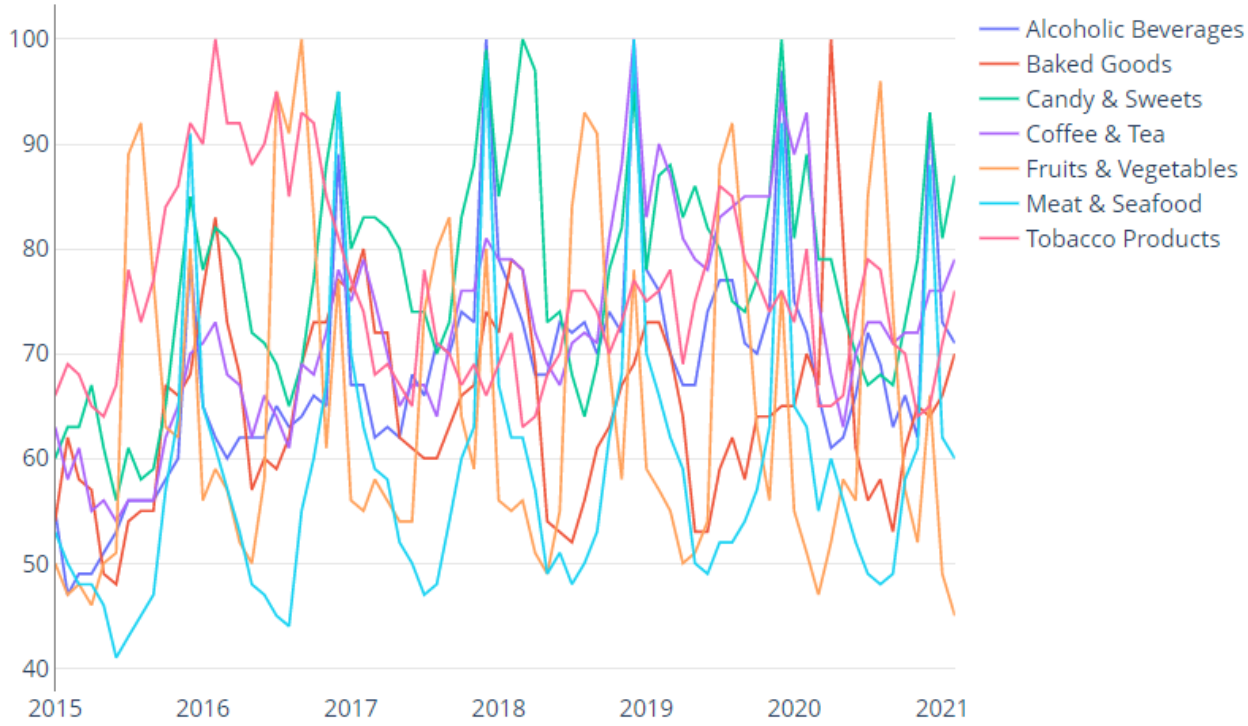


Рисунок 2 – Продовольственные категории Google Trends

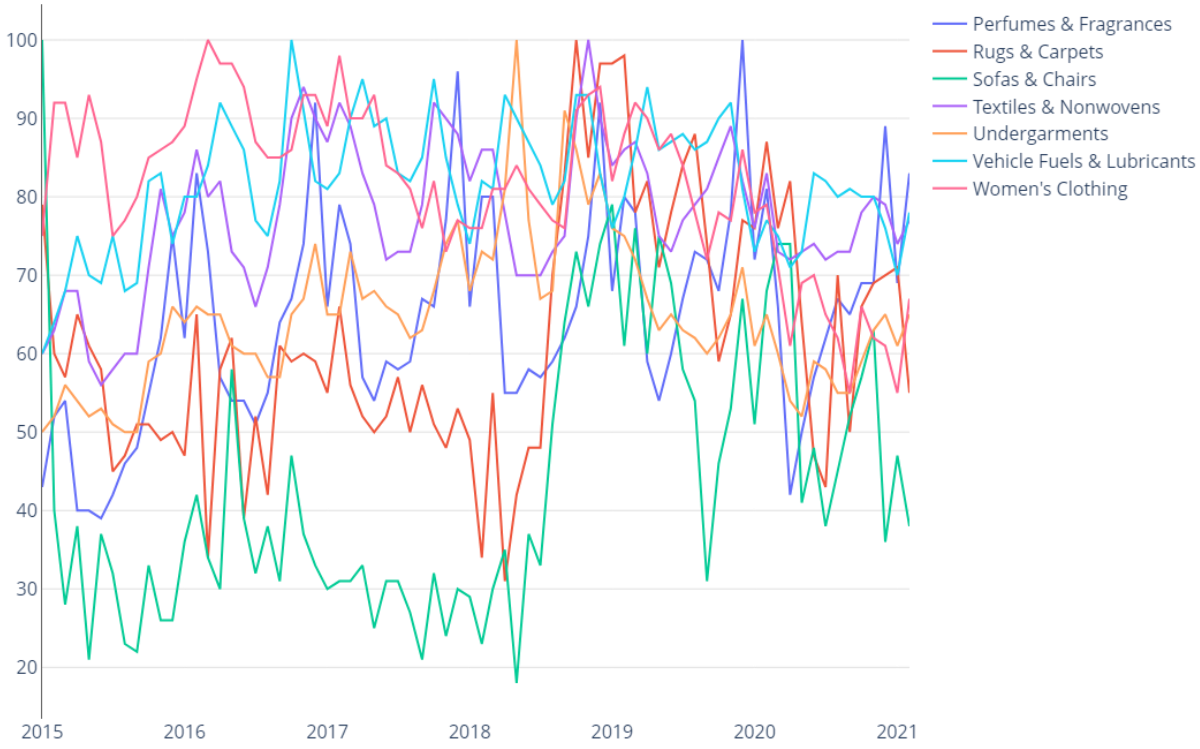


Рисунок 3 – Примеры непродовольственных категорий Google Trends

3 Среднесрочное прогнозирование с помощью Google Trends

В данном разделе мы оцениваем вклад, который вносят ряды Google Trends в прогнозирование продовольственных и непродовольственных товаров на несколько месяцев вперед. Чтобы учесть влияние предыдущих значений переменных на текущие показатели ретейла, мы используем запаздывающие значения макроэкономических переменных, в том

числе целевых. Рассматриваются до трех запаздываний включительно для каждой переменной. Далее, создаем базовые модели, которые будем пытаться улучшить путем добавления трендов (Google Trends). Для этого рассматриваем все возможные комбинации моделей из не более чем трех объясняющих переменных (без трендов). Для каждой комбинации применяется методология, изложенная ниже.

1. Нормируем все рассматриваемые переменные в диапазоне от 0 до 100 для унификации их измерений. Делим выборку на тренировочную и тестовую. В тестовую выборку включаются 10 последних значений всех переменных, что составляет 86% всей выборки. Если в регрессии участвуют лаги, то размер тренировочной выборки сокращается на их количество.

2. На тренировочной выборке оцениваем модель методом наименьших квадратов. Мы моделируем прогнозы в предположении, что будущие значения ретейла нам неизвестны. В моделях с лагированными значениями целевой переменной мы учитываем, что данные по ретейлу публикуются не моментально, а с задержкой приблизительно в один месяц. Поэтому на тестовой выборке, начиная со второго шага, реальное значение первого запаздывания ретейла заменяется его оценкой, полученной на предыдущем шаге. Аналогично, второй лаг заменяется оценкой, начиная со третьего шага, третий – с четвертого.

3. Строим прогноз на 10 наблюдений вперед и, используя данные тестовой выборки, считаем среднюю ошибку прогноза. В качестве метрики для ошибки мы рассматриваем среднюю абсолютную ошибку прогноза (MAE).

Из полученных комбинаций мы выбираем 10 (в целях транспарентности результатов) моделей с самым высоким R^2. Данные модели мы будем называть базовыми. В них будем добавлять различные комбинации трендов (не более двух) и смотреть, улучшаются ли значения коэффициента детерминации и ошибки прогноза. Уточним, что мы используем скорректированный R^2, так как это дает возможность сравнивать модели с разным количеством регрессоров. Затем мы повторяем методологию выше для каждой серии моделей. Наконец, сопоставляем полученные значения R^2 и MAE для регрессий с трендами и без них. Если коэффициент детерминации вырос, и наряду с этим снизилась ошибка прогноза, то мы считаем, что тренды улучшают прогнозы для ретейла.

В расчетах используются сокращенные обозначения для переменных. Целевые переменные обозначаются как food_goods и nonfood_goods для продовольственного и непродовольственного ретейлов соответственно. Среди объясняющих переменных ipd обозначает индекс потребительского доверия, real_rate – реальный валютный курс, oil – реальную цену на нефть и vix – индекс волатильности. Добавление «-1», «-2» и «-3» к обозначениям указывает на величину лага. В таблицах 1 и 2 представлены первые 10

комбинаций переменных в регрессиях для продовольственных и непродовольственных товаров соответственно с наибольшим R^2.

Таблица 1 – Примеры базовых и расширенных моделей для продовольственного ретейла

	Регрессоры	Значимые регрессоры	R^2	MAE
1	const, ipd-3, real_rate-1, food_goods-2	const, ipd-3, real_rate-1, food_goods-2	0.804	17.55
2	const, ipd-3, real_rate-1, food_goods-3	const, ipd-3, real_rate-1, food_goods-3	0.802	22.00
3	const, ipd-1, real_rate-1, food_goods-3	const, ipd-1, real_rate-1, food_goods-3	0.796	19.89
4	const, ipd-3, real_rate, food_goods-2	const, ipd-3, real_rate, food_goods-2	0.787	17.57
5	const, oil-1, food_goods-2	oil-1, food_goods-2	0.784	18.36
6	const, ipd-2, real_rate-1, food_goods-2	ipd-2, real_rate-1, food_goods-2	0.783	19.99
7	const, ipd-3, real_rate, food_goods-3	const, ipd-3, real_rate, food_goods-3	0.774	21.80
8	const, ipd-3, real_rate-2, food_goods-2	const, ipd-3, real_rate-2, food_goods-2	0.772	14.77
9	const, real_rate-1, food_goods-1	real_rate-1, food_goods-1	0.772	13.94
10	const, ipd-2, real_rate-1, food_goods-3	const, ipd-2, real_rate-1, food_goods-3	0.770	20.66

Примечательно, что в каждой регрессии встречаются лаги целевой переменной, и в большинстве регрессий присутствуют лаги индикатора потребительского доверия и валютного курса. Спрос на продовольственные товары менее волатилен, чем на непродовольственные, поэтому предыдущие значения ретейла хорошо объясняют текущие. Также ввиду значительного импорта среди продовольственных товаров (29% на конец 2020[6]) и существования закона единой цены для торгуемых товаров валютный курс также является фактором, объясняющим динамику продовольственного ретейла. Наконец, индикатор потребительского доверия коррелирует с рынком потребительских товаров, так как характеризует платежеспособность населения.

Таблица 2 – Примеры базовых и расширенных моделей для непродовольственного ретейла

	Регрессоры	Значимые регрессоры	R^2	MAE
1	const, ipd, oil-1	const, ipd, oil-1	0.649	11.69
2	const, oil, real_rate, real_rate-1	oil, real_rate, real_rate-1	0.635	7.07
3	const, ipd, oil	const, ipd, oil	0.616	11.70
4	const, ipd, real_rate-1	const, ipd, real_rate-1	0.614	11.01
5	const, ipd-1, oil-1	const, ipd-1, oil-1	0.589	12.44
6	const, ipd-1, oil	const, ipd-1, oil	0.588	9.98
7	const, ipd, real_rate-2	const, ipd, real_rate-2	0.585	10.71
8	const, ipd, vix-2	const, ipd, vix-2	0.566	14.00
9	const, ipd, real_rate	const, ipd, real_rate	0.564	13.16
10	const, oil-1	oil-1	0.563	11.94

[6] https://rosstat.gov.ru/folder/11188

Среди лучших регрессий для непродовольственных товаров не встречаются лаги целевой переменной. Текущее значение индикатора потребительского доверия, цену на нефть и их первые запаздывания можно заметить почти во всех рассматриваемых комбинациях переменных. Индикатор потребительского доверия отражает спрос на дорогие покупки, среди которых есть те, которые входят в непродовольственный ретейл. Например, покупка легкового автомобиля, мебели или мобильного телефона. Цены на нефть через канал валютного курса влияют на рынок непродовольственных товаров через закупки промежуточных товаров из-за рубежа или импорт готовых продуктов. Кроме того, шоки цены на нефть напрямую влияют на благосостояние населения, а, значит, и на платёжеспособность.

Далее, мы добавляем к полученным базовым моделям продовольственные и непродовольственные тренды соответственно. Для продовольственных товаров были подобраны 38 комбинаций, улучшающих прогнозы и повышающих долю объясняемой вариации продовольственного ретейла. В дальнейшем мы будем называть такие модели расширенными. Диаграмма 4 демонстрирует статистику среди трендов в расширенных моделях.

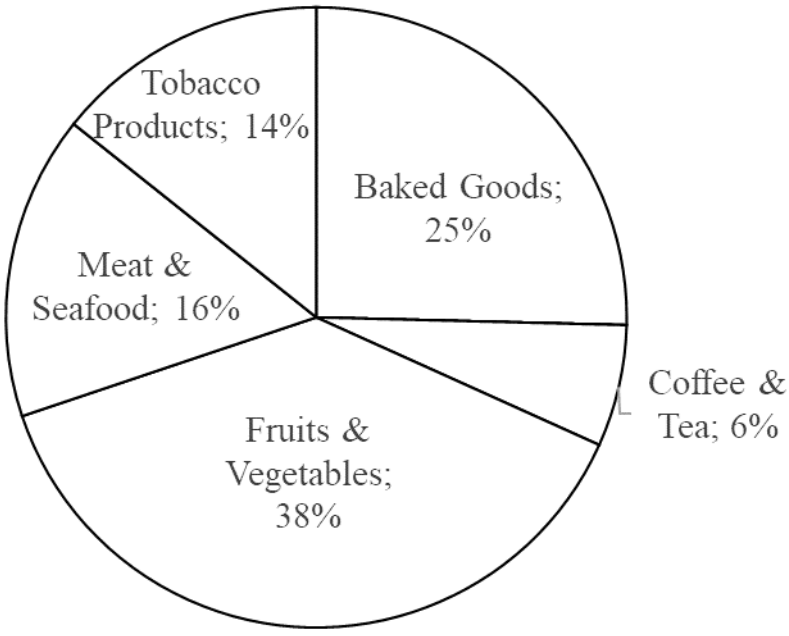


Рисунок 4 – Статистика по продовольственным трендам в расширенных моделях

В значительную часть моделей входят Fruits & Vegetables. Однако добавление данного тренда в модель не сильно поднимает значение R^2 и незначительно улучшает прогнозы. Напротив, для второго по частоте тренда Baked Goods можно сказать обратное. Также примечательно, что оба этих тренда чаще прочих оказываются значимыми после оценки коэффициентов моделей.

В таблице 3 указаны примеры моделей, в которых сильнее всего увеличивается R^2 при добавлении трендов. Примечательно, что всех моделях, по которым усреднялись прогнозы, присутствует тренд Baked Goods.

Таблица 3 – Примеры базовых и расширенных моделей для продовольственного ретейла

Регрессоры	Значимые регрессоры	Регрессоры (тренды)	Значимые регрессоры (тренды)	Разница MAE	Разница R^2
const, ipd-2, real_rate-1, food_goods-3	const, ipd-2, real_rate-1, food_goods-3	const, ipd-2, real_rate-1, food_goods-3, Baked Goods, Meat & Seafood	real_rate-1 food_goods-3 Baked Goods	-8.225	0.084
const, oil-1, food_goods-2	oil-1, food_goods-2	const, oil-1, food_goods-2, Baked Goods	const, food_goods-2, Baked Goods	-1.207	0.058
const, oil-1, food_goods-2	oil-1, food_goods-2	const, oil-1, food_goods-2, Fruits & Vegetables, Tobacco Products	food_goods-2, Fruits & Vegetables	-2.468	0.028
const, ipd-3, real_rate-2, food_goods-2	const, ipd-3, real_rate-2, food_goods-2	const, ipd-3, real_rate-2, food_goods-2, Fruits & Vegetables, Tobacco Products	const, ipd-3, real_rate-2, food_goods-2, Fruits & Vegetables	-1.488	0.018

Для непродовольственных товаров получилось 218 комбинаций, улучшающих прогнозы и повышающих R^2. На рисунке 5 представлена статистика среди основных трендов, которые вошли в эти комбинации.

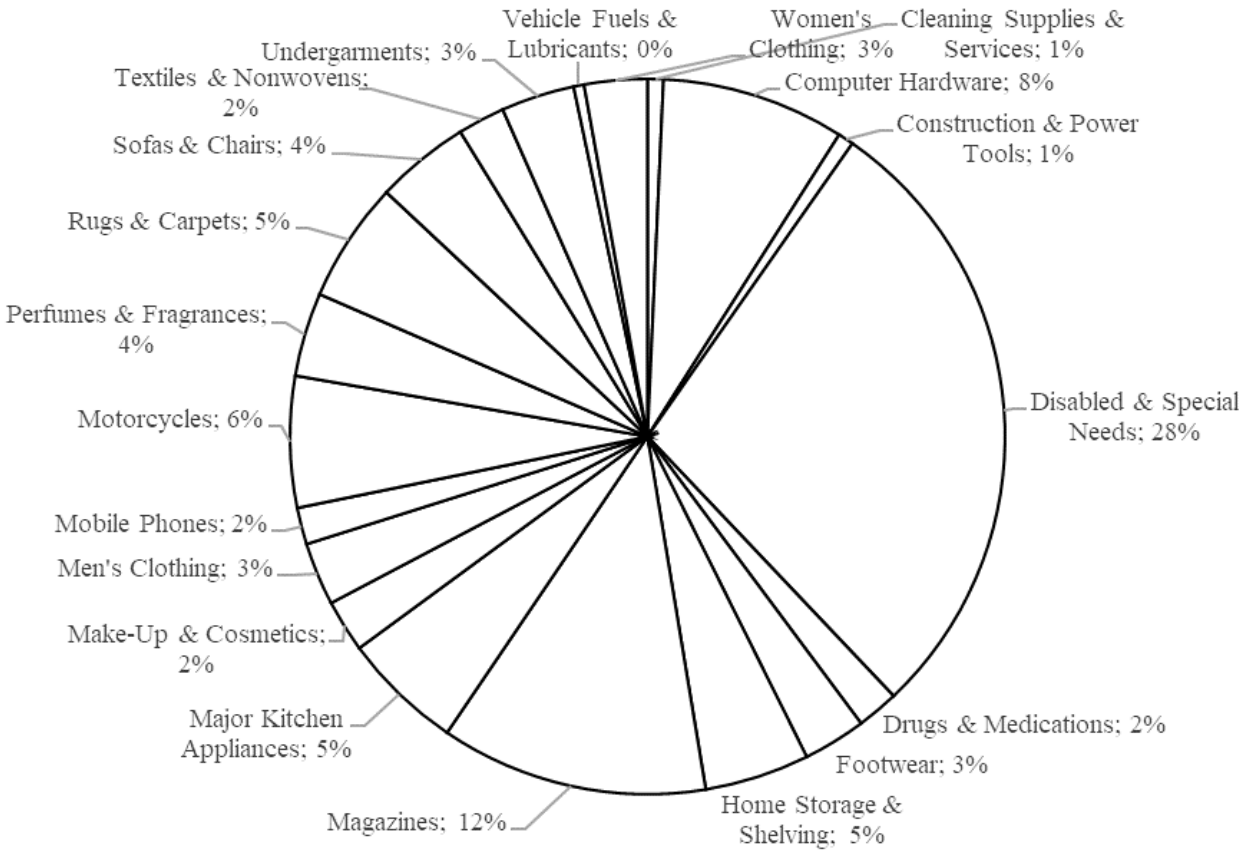


Рисунок 5 – Статистика по непродовольственным трендам в расширенных моделях

В таблице 4 указаны примеры моделей, улучшенных трендами. Интересы к товарам для благоустройства помещений (Rugs & Carpets, Sofas & Chairs, Home Storage & Shelving), чистящим средствам и услугам по уборке (Cleaning Supplies & Services), а также к журналам и газетам (Magazines) лучше всех помогают предсказывать непродовольственный ретейл. Примечательно, что непродовольственные тренды помогают повысить долю объяснённой вариации зависимой переменной практически на четверть, тогда как улучшения результатов для R^2 среди продовольственных товаров не достигают и 10 п.п.

Таблица 4 – Примеры базовых и расширенных моделей для непродовольственного ретейла

Регрессоры	Значимые регрессоры	Регрессоры (тренды)	Значимые регрессоры (тренды)	Разница MAE	Разница R^2
const, ipd, real_rate	const, ipd, real_rate	const, ipd, real_rate, Magazines, Rugs & Carpets	const, ipd, Magazines, Rugs & Carpets	-2.063	0.206
const, ipd-1, oil	const, ipd-1, oil	const, ipd-1, oil, Cleaning Supplies & Services, Magazines	const, ipd-1, oil, Cleaning Supplies & Services, Magazines	-1.827	0.198
const, ipd, real_rate	const, ipd, real_rate	const, ipd, real_rate, Magazines, Sofas & Chairs	const, ipd, Magazines, Sofas & Chairs	-1.458	0.189

const, ipd, real_rate-2	const, ipd, real_rate-2	const, ipd, real_rate-2, Home Storage & Shelving, Magazines	const, ipd, Home Storage & Shelving	-2.397	0.179

4 Псевдовневыборочный наукастинг агрегированного ретейла с помощью Google Trends

Ввиду того, что данные по ретейлу выходят с задержкой примерно в 1 месяц, часто стоит задача прогнозирования данного показателя лишь на месяц вперёд, например, как в работе [22]. В данном разделе мы оцениваем вклад, который вносят ряды Google Trends в прогнозирование продовольственных и непродовольственных товаров с использованием псевдовневыборочного наукастинг. Для унификации результатов в целях удобства их отображения на графиках мы нормируем все переменные в диапазоне от 0 до 100. Подобная нормировка не умаляет качества прогнозных моделей, однако позволит сопоставить ошибки прогнозирования разных ретейлов. Чтобы учесть влияние предыдущих значений переменных на текущие показатели ретейла, мы используем лагированные значения макроэкономических переменных, в том числе целевых. Рассматриваются до трех запаздываний включительно для каждой переменной. Далее, создаем базовые модели, которые будем пытаться улучшить путем добавления трендов (Google Trends). Для этого рассматриваем все возможные комбинации моделей из не более чем трех[7] объясняющих переменных (без трендов). Для каждой комбинации применяется методология, изложенная ниже.

1. Делим выборку на тренировочную и тестовую. В тестовую выборку включаются 40% последних значений всех переменных. Если в регрессии участвуют лаги, то размер тестовой выборки пропорционально сокращается.

2. На тренировочной выборке оцениваем модель методом наименьших квадратов.

3. Строим прогноз на тестовой выборке на один шаг вперед, получаем оценку целевой переменной. Добавляем это наблюдение в тренировочную выборку, снова оцениваем модель с учетом нового наблюдения. Данную процедуру продолжаем до конца тестовой выборки. В конце считаем среднюю абсолютную ошибку прогноза, используя оцененные значения целевой переменной. В качестве метрики для ошибки мы рассматриваем среднюю абсолютную ошибку прогноза (MAE).

Из полученных комбинаций мы выбираем 10 (в целях транспарентности результатов) моделей с самым высоким R^2. Данные модели мы будем называть базовыми. В них будем

7 Мы также пробовали использовать наборы переменных с большим количеством лагов (до 6), но они практически никогда не входили базовые модели.

добавлять различные комбинации трендов (не более двух) и смотреть, улучшаются ли значения коэффициента детерминации и ошибки прогноза. Уточним, что мы используем скорректированный R^2, так как это дает возможность сравнивать модели с разным количеством регрессоров. Затем мы повторяем методологию выше для каждой серии моделей. Наконец, сопоставляем полученные значения R^2 и MAE для регрессий с трендами и без них. Если коэффициент детерминации вырос, и наряду с этим снизилась ошибка прогноза, то мы считаем, что тренды улучшают прогнозы для ретейла.

В расчетах используются сокращенные обозначения для переменных. Целевые переменные обозначаются как food_goods и nonfood_goods для продовольственного и непродовольственного ретейлов соответственно. Среди объясняющих переменных ipd обозначает индекс потребительского доверия, real_rate – реальный валютный курс, oil – реальную цену на нефть и rvi – индекс волатильности. Добавление «-1», «-2» и «-3» к обозначениям указывает на величину лага.

Уточним, что мы ранжируем таблицы по R^2, полученному из последней оцененной модели в ходе прогнозов. Таким образом, данный R^2 показывает долю объясненной вариации всей рассматриваемой выборки без последнего наблюдения. В таблицах 5 и 6 указаны лучшие модели для продовольственных и непродовольственных товаров.

Таблица 5 – Базовые модели для продовольственного ретейла

	Регрессоры	Значимые регрессоры	R^2	MAE
1	const, real_rate-1, food_goods-1	real_rate-1, food_goods-1	0.81	12.58
2	const, real_rate, food_goods-1	real_rate, food_goods-1	0.80	12.60
3	const, real_rate-2, food_goods-1	real_rate-2, food_goods-1	0.79	12.82
4	const, food_goods-1	food_goods-1	0.78	12.84
5	const, real_rate-3, food_goods-1	real_rate-3, food_goods-1	0.78	13.16
6	const, ipd-1, real_rate-1, food_goods-2	const, ipd-1, real_rate-1, food_goods-2	0.77	14.75
7	const, ipd-1, real_rate-2, food_goods-2	const, ipd-1, real_rate-2, food_goods-2	0.76	15.06
8	const, ipd-1, real_rate-1, food_goods-3	const, ipd-1, real_rate-1, food_goods-3	0.76	14.19
9	const, ipd-1, real_rate-2, food_goods-3	const, ipd-1, real_rate-2, food_goods-3	0.76	14.70
10	const, ipd, real_rate-1, food_goods-3	const, ipd, real_rate-1, food_goods-3	0.75	13.28

Примечательно, что в каждой регрессии встречаются лаги целевой переменной, и в значительной части регрессий присутствуют лаги индикатора потребительского доверия и валютного курса. Спрос на продовольственные товары менее волатилен, чем на непродовольственные, поэтому предыдущие значения ретейла хорошо объясняют текущие. Также ввиду значительного импорта среди продовольственных товаров (29% на конец 20208) и существования закона единой цены для торгуемых товаров валютный курс также является

8 https://rosstat.gov.ru/folder/11188

фактором, объясняющим динамику продовольственного ретейла. Наконец, индикатор потребительского доверия коррелирует с рынком потребительских товаров, так как характеризует платежеспособность населения.

Таблица 6 – Примеры базовых моделей для непродовольственного ретейла

	Регрессоры	Значимые регрессоры	R^2	MAE
1	const, oil, real_rate, real_rate-1	const, oil, real_rate, real_rate-1	0.59	13.65
2	const, oil, rvi, rvi-1	const, oil, rvi, rvi-1	0.58	13.32
3	const, ipd, rvi, rvi-1	const, ipd, rvi, rvi-1	0.58	13.31
4	const, oil-1, real_rate, real_rate-1	const, oil-1, real_rate, real_rate-1	0.56	14.90
5	const, ipd-1, oil-1	const, ipd-1, oil-1	0.56	14.41
6	const, ipd, real_rate, real_rate-1	const, ipd, real_rate, real_rate-1	0.56	12.40
7	const, ipd-1, rvi, rvi-1	const, ipd-1, rvi, rvi-1	0.55	14.56
8	const, ipd, real_rate-1	const, ipd, real_rate-1	0.53	12.67
9	const, oil, real_rate, real_rate-2	const, oil, real_rate, real_rate-2	0.53	13.95
10	const, ipd, oil	const, ipd, oil	0.53	12.30

Среди лучших регрессий для непродовольственных товаров не встречаются лаги целевой переменной. Текущее значение индикатора потребительского доверия, цену на нефть и их первые запаздывания можно заметить почти во всех рассматриваемых комбинациях переменных. Индикатор потребительского доверия отражает спрос на дорогие покупки, среди которых есть те, которые входят в непродовольственный ретейл. Например, покупка легкового автомобиля, мебели или мобильного телефона. Цены на нефть посредством канала валютного курса влияют на рынок непродовольственных товаров через закупки промежуточных товаров из-за рубежа, импорт готовых продуктов. Кроме того, шоки цены на нефть напрямую влияют на благосостояние населения, а, значит, и на платёжеспособность.

При добавлении трендов в регрессии с продовольственными товарами получилось 80 расширенных комбинаций. Статистика по популярности трендов в полученных моделях приведена на рисунке 6.

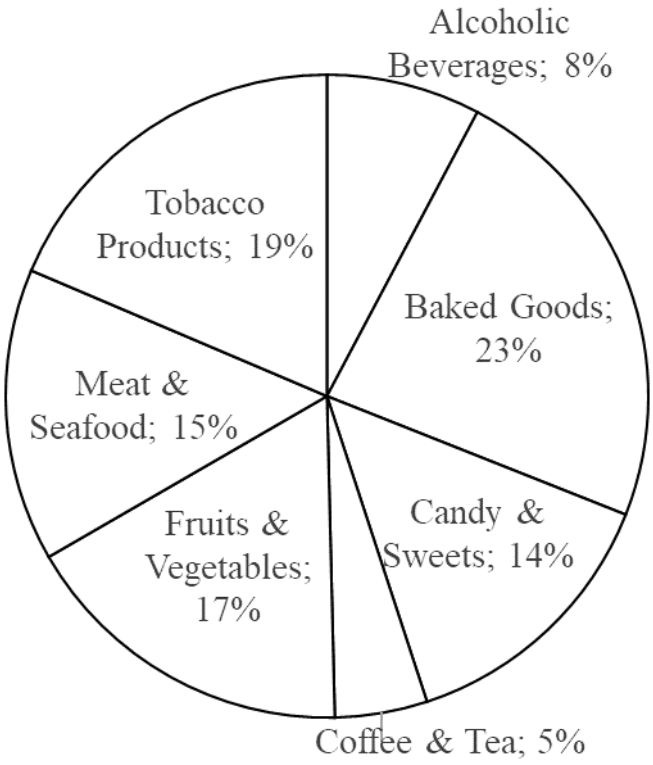


Рисунок 6 – Статистика по продовольственным трендам в расширенных моделях

Некоторые из расширенных регрессий представлены в таблице 7. Самый большой вклад в повышение доли объясняемой вариации вносит тренд Baked Goods, также в значительной части регрессий встречается тренд Meat & Seafood. Прочие тренды незначительно снижают ошибку прогноза и улучшают значение R^2.

Таблица 7 – Примеры базовых и расширенных моделей для продовольственного ретейла

Регрессоры	Значимые регрессоры	Регрессоры (тренды)	Значимые регрессоры (тренды)	Разница MAE	Разница R^2
const, ipd-1, real_rate-2, food_goods-3	const, ipd-1, real_rate-2, food_goods-3	const, ipd-1, real_rate-2, food_goods-3, Baked Goods, Tobacco Products	real_rate-2, food_goods-3, Baked Goods, Tobacco Products	-5.65	0.12
const, ipd-1, real_rate-2, food_goods-3	const, ipd-1, real_rate-2, food_goods-3	const, ipd-1, real_rate-2, food_goods-3, Baked Goods, Coffee & Tea	const, real_rate-2, food_goods-3, Baked Goods, Coffee & Tea	-4.55	0.10
const, ipd, real_rate-1, food_goods-3	const, ipd, real_rate-1, food_goods-3	const, ipd, real_rate-1, food_goods-3, Meat & Seafood	ipd, real_rate-1, food_goods-3, Meat & Seafood	-1.13	0.05
const, ipd, real_rate-1, food_goods-3	const, ipd, real_rate-1, food_goods-3	const, ipd-1, rvi-3, real_rate-2, Fruits & Vegetables	const, ipd-1, rvi-3, real_rate-2, Fruits & Vegetables	-0.12	0.03

Для непродовольственных трендов было получено, что в 206 регрессиях тренды улучшают коэффициент детерминации и снижают ошибку прогноза. На рисунке 7 приведена

статистика по популярности трендов среди полученных комбинаций. Во всех моделях, улучшающих ошибку прогноза сильнее всех, встречается тренд Magazines.

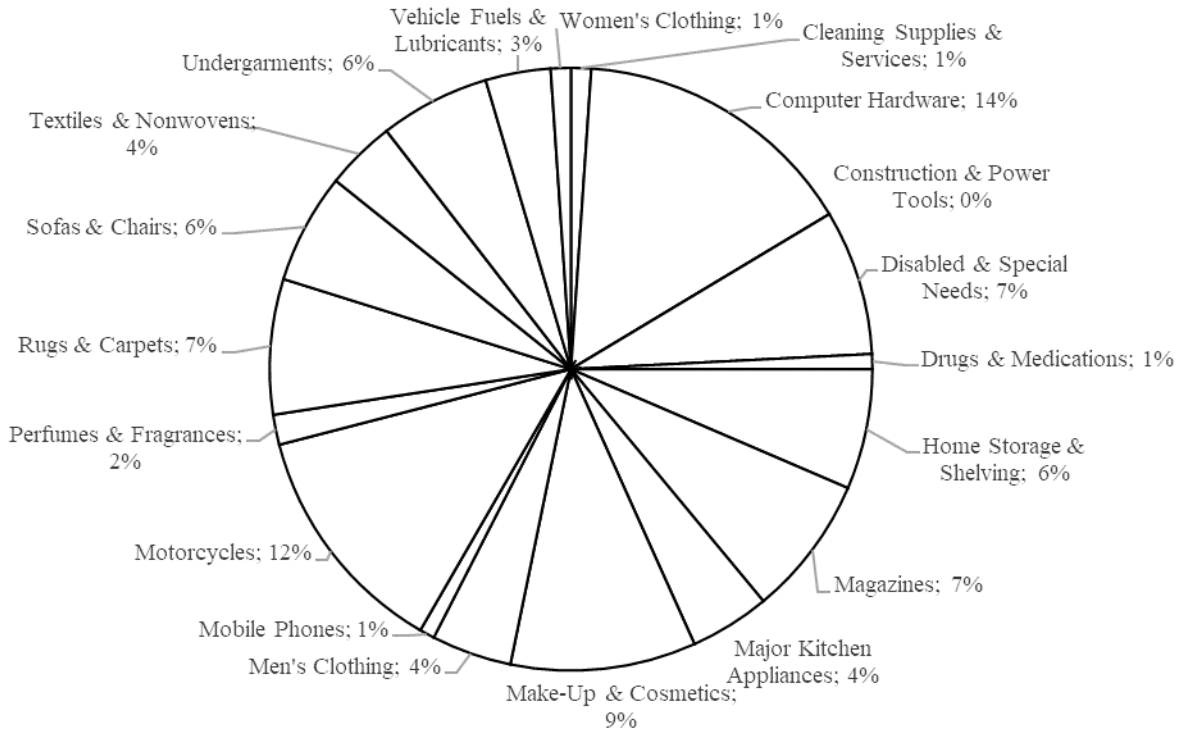


Рисунок 7 – Статистика по непродовольственным трендам в расширенных моделях

В таблице 8 представлены некоторые из регрессий с добавлением Google Trends. Среди трендов, максимально увеличивающих коэффициент детерминации, встречаются Magazines, а также Motorcycles и товары по благоустройству дома. Также можно наблюдать, что тренд Computer Hardware значительно помогает повысить долю объясненной вариации непродовольственного ретейла на рассматриваемом промежутке.

Таблица 8 – Примеры базовых и расширенных моделей для непродовольственного ретейла

Регрессоры	Значимые регрессоры	Регрессоры (тренды)	Значимые регрессоры (тренды)	Разница MAE	Разница R^2
const, ipd-1, rvi, rvi-1	const, ipd-1, rvi, rvi-1	const, ipd-1, rvi, rvi-1, Magazines, Rugs & Carpets	const, ipd-1, rvi-1, Magazines, Rugs & Carpets	-3.09	0.17
const, ipd-1, oil-1	const, ipd-1, oil-1	const, ipd-1, oil-1, Magazines, Sofas & Chairs	const, ipd-1, oil-1, Magazines, Sofas & Chairs	-2.61	0.16
const, oil, real_rate, real_rate-2	const, oil, real_rate, real_rate-2	const, oil, real_rate, real_rate-2, Home Storage & Shelving	const, oil, real_rate-1, Home Storage & Shelving, Magazines	-1.16	0.14
const, ipd, real_rate-1	const, ipd, real_rate-1	const, ipd, real_rate-1, Computer Hardware, Motorcycles	const, ipd, real_rate-1, Computer Hardware	-0.89	0.10

ЗАКЛЮЧЕНИЕ

В данной работе на основе анализа эмпирической литературы мы предприняли попытку понять, насколько может быть полезна информация, содержащаяся в Google Trends, для прогнозирования различных экономических показателей. В результате мы обнаружили, что использование подобных запросов является популярным инструментом, помогающим улучшить качество прогнозов таких разнообразных показателей, как розничные продажи, продажи автомобилей, домов, туристических продуктов, а также расходы домашних хозяйств на потребление.

В качестве примера использования Google Trends для увеличения прогнозной силы моделей, мы рассмотрели поисковые запросы по различным товарным группам и оценили вклад, который они вносят в прогнозную силу моделей для продовольственных и непродовольственных розничных продаж. Мы использовали модели линейной регрессии для среднесрочного и краткосрочного прогнозирования ретейла с добавлением трендов.

Для базовых моделей были использованы такие макроэкономические переменные, как российский индекс волатильности, цена на нефть марки Brent, реальный эффективный валютный курс и индикатор потребительского доверия от ВЦИОМ. В ходе работы мы старались улучшить прогнозы по базовым моделям путем добавления отдельных трендов.

Было выявлено, что на рассмотренном промежутке при среднесрочном прогнозировании тренды Baked Goods и Fruits & Vegetables сильнее прочих повышают коэффициент детерминации регрессии для продовольственного ретейла, улучшая при этом среднюю ошибку прогноза. В регрессиях с непродовольственными товарами наибольшее улучшение R^2 происходит за счет тренда Magazines, а также трендов, связанных с благоустройством помещений (Rugs & Carpets, Sofas & Chairs, Home Storage & Shelving).

При краткосрочном прогнозировании для продовольственных товаров тренды Baked Goods и Meat & Seafood оказались лучшими в расширенных моделях. В случае с непродовольственным ретейлом прогнозы улучшаются путем добавления трендов Magazines и Computer Hardware, и также трендов, связанных с товарами для дома.

Примечательно, что существенная часть тестовой выборки пересекается с периодом сильного шока глобального спроса, вызванного пандемией, что добавляет значимости полученным результатам.

В качестве направлений для дальнейших исследований может рассматриваться использование трендов как для прогнозирования совокупного потребления в российской экономике, так и для прогнозирования розничных продаж конкретных российских ретейлеров. Полученные в данной работе результаты могут быть полезны как профильным органам, ответственным за соответствующие направления экономической политики, так и частным ретейлерам для повышения эффективности бизнеса.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Реальные располагаемые доходы россиян рекордно упали из-за пандемии // РБК. 2020.
2. Choi H., Varian H. Predicting the Present with Google Trends // Google Technical Report. 2009.
3. Vossen S., Schmidt S. Forecasting private consumption: Survey-based indicators vs. Google Trends // Journal of Forecasting. 2011. Vol. 30. No. 6. pp. 565–578.
4. Boone T., Ganeshan R., Hicks R. L., Sanders N. R. Can Google Trends Improve Your Sales Forecast? // Production and Operations Management Society. 2017. Vol. 27. No. 10. pp. 1770–1774.
5. Kuchler F., Bowman M., Sweitzer M., Greene C. Evidence from Retail Food Markets That Consumers Are Confused by Natural and Organic Food Labels // Journal of Consumer Policy. 2020. No. 43. pp. 379 – 395.
6. Silva E. S., Hassani H., Madsen D. Ø., Gee L. Googling Fashion: Forecasting Fashion Consumer Behaviour Using Google Trends // Social sciences. 2019. Vol. 8. No. 111. pp. 1-23.
7. Oh J., Ha K-J., Jo Y-H. New Normal Weather Breaks a Traditional Clothing Retail Calendar // Preprint from Research Square. 2021. pp. 1-19.
8. Ellingsen J. Can Google search indices nowcast Norwegian retail sales and unemployment rate? // University of Oslo Library. 2017. pp. 1-54.
9. Wu L., Brynjolfsson E. The future of prediction: How Google searches foreshadow housing prices and sales. // Economic Analysis of the Digital Economy. 2015. pp. 89–118.
10. Dietzel A.M. Sentiment-based predictions of housing market turning points with Google trends // International Journal of Housing Markets and Analysis. 2016. Vol. 9. No. 1. pp. 108 - 136.
11. Wijnhoven F., Plant O. Sentiment Analysis and Google Trends Data for Predicting Car Sales // Thirty Eighth International Conference on Information Systems. 2017. pp. 1-16.
12. Li X., Pan B., Law R., Huang X. Forecasting tourism demand with composite search index // Tourism Management. 2017. No. 59. pp. 57-66.

13. Padhi S. S., Pati R. K. Quantifying potential tourist behavior in choice of destination using Google Trends // Tourism Management Perspectives. 2017. No. 24. pp. 34-47.

14. Dergiades T. M.E..P.B. Google Trends and tourists' arrivals: Emerging biases and proposed corrections // Tourism Management. 2018. No. 66. pp. 108-120.

15. Tijerina J. D., Morrison S. D., Nolan I. T., Parham M. J. Predicting Public Interest in Nonsurgical Cosmetic Procedures Using Google Trends // The American Society for Aesthetic Plastic Surgery, Inc. 2019. pp. 1-30.

16. Fasulo A., Terribili M.D., Guandalini A. Google Trends for Nowcasting Quarterly Household Consumption Expenditure // Rivista Italiana di Economia, Demografia e Statistica. 2018. Vol. 71. No. 4. pp. 5-13.

17. Scott S. L., Varian H. R. Bayesian Variable Selection for Nowcasting Economic Time Series // National Bureau of Economic Research. 2015. pp. 119-135.

18. Woo J., Owen A.L. Forecasting private consumption with Google Trends data // Journal of Forecasting. 2019. No. 38. pp. 81–91.

19. Pekar V. Purchase Intentions on Social Media as Predictors of Consumer Spending // Proceedings of the International AAAI Conference on Web and Social Media. 2020. Vol. 14. No. 1. pp. 545-556.

20. Kilian L. Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market // The American Economic Review. 2009. Vol. 99. No. 3. pp. 1053-1069.

21. Polbin A., Skrobotov A., Zubarev A. How the oil price and other factors of real exchange rate dynamics affect real GDP in Russia // Emerging Markets Finance and Trade. 2020. Vol. 56. No. 15. pp. 3732-3745.

22. Майорова К., Фокин Н. Наукастинг темпов роста стоимостных объемов экспорта и импорта России по товарным группам // Деньги и Кредит. 2021. Vol. 80. No. 3. pp. 34-48.

23. Bureau U.S.C. X-13ARIMA-SEATS 2013.

Приложение А. Результаты теста Дики-Фуллера на наличие единичных корней в Google Trends

Таблица А.1 – p-значения для продовольственных категорий Google Trends

Название переменной	Alcoholic Beverages	Baked Goods	Candy & Sweets	Coffee & Tea	Fruits & Vegetables	Meat & Seafood	Tobacco products
p-значение	0.03	0.02	0.14	0.34	0.3	0.07	0.0

Примечание – ряды предварительно были очищены от сезонности с помощью процедуры X13ARIMA-SEATS [25]. Нулевая гипотеза о нестационарности ряда отвергается при p-значении < 0.05.

Таблица А.2 – p-значения для непродовольственных категорий Google Trends

Название переменной	1	2	3	4	5	6	7	8	9	10	11	12	13
p-значение	0.47	0.1	0.36	0.14	0.04	0.4	0.82	0.1	0.09	0.01	0.13	0.17	0.0

Продолжение таблицы А.2

Название переменной	14	15	16	17	18	19	20	21	22	23	24	25
p-значение	0.06	0.39	0.1	0.57	0.85	0.27	0.11	0.11	0.01	0.34	0.11	0.74

Примечание: 1 - Audio Equipment, 2 - Bicycles & Accessories, 3 - Book Retailers, 4 - Children's Clothing, 5 - Classic Vehicles, 6 - Cleaning Supplies & Services, 7 - Computer Hardware, 8 - Construction & Power Tools, 9 - Disabled & Special Needs, 10 - Drugs & Medications, 11 – Footwear, 12 - Home Storage & Shelving, 13 – Magazines, 14 - Major Kitchen Appliances, 15 - Make-Up & Cosmetics, 16 - Men's Clothing, 17 - Mobile Phones, 18 – Motorcycles, 19 - Perfumes & Fragrances, 20 - Rugs & Carpets, 21 - Sofas & Chairs, 22 - Textiles & Nonwovens, 23 – Undergarments, 24 - Vehicle Fuels & Lubricants, 25 - Women's Clothing. Ряды предварительно были очищены от сезонности с помощью процедуры X13ARIMA-SEATS [25]. Нулевая гипотеза о нестационарности ряда отвергается при p-значении < 0.05.

Приложение Б. Ссылки на данные по макроэкономическим переменным

Выручка с совокупных розничных продаж	https://rosstat.gov.ru/folder/23457
Индекс волатильности	https://ru.investing.com/indices/russian-vix
Индикатор потребительского доверия	https://wciom.ru/ratings/indeks-potrebitelskogo-doverija
Нефть марки Brent	https://fred.stlouisfed.org/series/POILBREUSDM
Реальный эффективный девизный валютный курс	https://fred.stlouisfed.org/series/RBRUBIS
Американский индекс потребительских цен	https://fred.stlouisfed.org/series/CPIAUCSL