Секция «Макроэкономика»

Наукастинг основных показателей российской экономики с помощью методов машинного обучения

Научный руководитель – Полбин Андрей Владимирович

Гареев Михаил Юрьевич

Студент (магистр)

Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Экономический факультет, Москва, Россия

E-mail: mkhlgrv@gmail.com

Работа посвящена наукастингу квартальных изменений основных макроэкономических показателей российской экономики - ВВП, потребление домохозяйств, инвестиции, экспорт, импорт, а также экспорт и импорт в долларах, с использованием большого набора данных (44 предиктора), включающих макроэкономическую статистику, показатели денежного и фондового рынка, опросные данные, цены на основные экспортные товары России и другие показатели, характеризующие внешние и внутренние факторы, влияющие на российскую экономику.

Для корректного наукастинга все данные были отсортированы в соответствии с датами публикации. Для науката на конец каждой недели квартала использовались данные, известные в этот период времени. Таким образом, с приближением к дате публикации переменной, информация, доступная для наукастинга или бэккастинга, росла.

Наукастинг представляет особый практический интерес, так как уже реализованные или реализовывающиеся в текущий момент времени показатели можно предсказать с достаточно большой точностью - многие обновляемые в реальном времени показатели содержат информацию о тех переменных, точные значения которых могут быть получены только после продолжительного лага, связанного со сбором статистики, а, благодаря тому, что предсказания могут быть получены задолго до публикации, их можно использовать - как властям, проводящим экономическую политику, так и частным лицам, принимающим экономические решения.

Для извлечения полезной информации из большого набора данных выбраны ансамблевые методы машинного обучения - бустинг и случайный лес, а также метод эластичной сети. Методы машинного обучения, в отличии от методов традиционной эконометрики, позволяют теоретически использовать неограниченное количество предикторов. Такой подход позволяет не заниматься отбором нескольких предикторов для прогноза, а использовать все доступные данные, чтобы получить из них информацию о ненаблюдаемых факторах, оказывающих влияние на экономику.

Для получения прогнозов в псевдореальном времени использовалось оценивание на скользящем расширяющемся окне. Выборка квартальных данных с 2-го квартала 2007 г. (левая граница связана с ограничениями, возникающими из-за включения некоторых коротких рядов, а также приведением рядов к стационарному виду) по 4-ый квартал 2020 г. была разбита на две части - тренировочную и тестовую. Для получения прогнозов на 1-ый квартал 2015 г. модель обучалась на данных, относящихся к периоду со 2-го квартала 2007 г по 4-ый квартал 2014 г. После построения прогнозов тренировочное окно сдвигалось на один квартал вперед, модель обучалась заново на новых данных, строился прогноз на следующий квартал тестовой выборки и так далее. Таким образом, для каждой недели квартала тестовой выборки создавался наукаст.

После получения прогнозов по ним была рассчитана метрика качества - RMSE. В качестве бенчмарка в работе выступают модели случайного блуждания и авторегрессии.

В результате эксперимента по прогнозированию было получено, что все переменные в среднем могут быть спрогнозированы с помощью методов машинного обучения лучше (по крайней мере, в смысле качества прогноза, посчитанного с помощью RMSE), чем с помощью случайного блуждания и авторегрессии, примерно с 10-ой недели квартала. Публикация ценовых индексов, а затем и показателей реального сектора (индекс промышленного производства, безработица и др.) позволяет, начиная с 10-ой недели квартала, улучшать качество предсказаний по сравнению с наивным прогнозом от 10% до 30% для разных переменных. В частности, хорошей иллюстрацией изменения науката по мере поступления новостей является 2-ой квартал 2020 г. Так, например, оценка методом эластичной сети изменения ВВП в процентах относительно 2-го квартала 2019 г. двигалась с 2% на первой неделе квартала до -5% на 23-ей неделе относительно начала квартала, а реальное изменение составило -8,5%. Дополнительный эксперимент с исключением почти всех высокочастотных финансовых данных из выборки, кроме цены на нефть и курса рубля к доллару, показал, что, несмотря на незначительное ухудшение качества предсказаний в первые недели квартала (до публикации релевантной статистики цен и реального сектора), ближе к концу квартала качество предсказаний моделей без использования финансовых данных оказывается не хуже, чем с их использованием.

Проведенные тесты на монотонность почти для всех переменных демонстрируют значимость следующих закономерностей: с увеличением доступной информации средняя квадратичная ошибка не увеличивается, а ковариация между прогнозами и истинными значениями не падает. Кроме того, для большинства переменных прогнозы оказываются несмещенными.