УДК 332.1+338.2
ББК 65.05+ 65.2/4
Э40

DOI 10.47711/978-5-907673-23-6

Федеральное государственное бюджетное учреждение науки Институт народнохозяйственного прогнозирования Российской академии наук

Федеральное государственное бюджетное учреждение науки Институт экономики и организации промышленного производства сибирского отделения Российской академии наук

Ответственные редакторы:
д-р экон. наук *А.А. Широв,*
д-р экон. наук *А.О. Баранов*

Э40 **Экономическая политика России в межотраслевом и пространственном измерении:** материалы конференции ИНП РАН и ИЭОПП СО РАН по межотраслевому и региональному анализу и прогнозированию (Россия, Московская область, 22-24 марта 2023 г.). – Т. 5 / отв. ред. А.А. Широв, А.О. Баранов. – Москва: НАУКА, 2023. – 176 с.

DOI 10.47711/978-5-907673-23-6
ISBN 978-5-907673-23-6

В книге представлены материалы пятой совместной конференции ИНП РАН и ИЭОПП СО РАН по межотраслевому и региональному анализу и прогнозированию, которая состоялась в г. Пересвет Московской области 22-24 марта 2023 г. В них представлен макроструктурный, пространственный и отраслевой подходы к анализу и прогнозированию социально-экономического развития России.

Для макроэкономистов, работников государственных органов власти, региональных властей и бизнеса, преподавателей, аспирантов, а также для читателей, интересующихся современными проблемами социально-экономического развития России.

УДК 332.1+338.2
ББК 65.05+ 65.2/4

ISBN 978-5-907673-23-6

5. Выпускники среднего профессионального образования на российском рынке труда: докл. к XXIV Ясинской (Апрельской) междунар. науч. конф. по проблемам развития экономики и общества, Москва, 2023 г. / Науч. ред. С. Ю. Рощин; М.: Изд. дом Высшей школы экономики, 2023. – 146 с.

***

Костин А.В., Родионова Д.А.

ФОРМИРОВАНИЕ СИСТЕМЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ НА ОСНОВЕ БАЗЫ ЗНАНИЙ ИЭОПП СО РАН[1]

Для решения задач анализа социально-экономического развития Азиатской России на основе синергии транспортной доступности и системных знаний о природно-ресурсном и промышленном потенциалах возникает необходимость расширения традиционного инструментария и создания единой Базы Знаний (БЗ). Такая БЗ разрабатывается в Институт экономики и организации промышленного производства Сибирского отделения Российской академии наук.

Создание БЗ проходит в несколько этапов:

- определение целей и задач её формирования,
- выстраивания архитектуры БЗ,
- создания Базы Данных (БД) и настройка её автоматического пополнения,
- формирование инструментария и модельного аппарата, для создания новых знаний с последующим их сохранением в системы БЗ,
- разработка веб-интерфейса.

Последующее расширение БЗ проходит в рамках развития последних трех направлений. Текущее наполнение БД является интенсивным и возникают проблемы, связанные с обработкой больших данных. Особенно это касается данных о компаниях, которые содержат

[1] *Работа выполнена по результатам исследования, проводимого при финансовой поддержке Министерства науки и высшего образования России в рамках крупного научного проекта «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий». Соглашение № 075-15-2020-804 от 02.10.2020 (грант № 13.1902.21.0016).*

информацию о 14.5 млн. компаний, включая бухгалтерскую отчетность, ОКВЭД и Гис-координаты. В ходе анализа данных, были выявлены ошибки и неточности, которые искажают результаты их аналитики.

В ходе интервью с экспертами было выявлено, что основной ошибкой при работе с данными из используемых нами источников при формировании базы данных компаний был резкий прирост выручки более чем в 100 или даже в 1000 раз за год. Также в ходе разведывательного анализа данных был выявлен ряд ошибок в других показателях отчётности компании. Предположительно, данные ошибки появились в следствии некорректно внесённой информации первоисточником. Для выявления ошибок, был разработан “Модуль по оценки уровня достоверности данных о финансовой отчётности компаний”, в рамках которого находятся компании с темпом прироста выручки превышающем 1000 раз и отклонением показатели отчётности от агрегированных показателей.

Были сформированы следующие этапы разработки программного модуля для реализации метода проверки темпов прироста компании:

- Загрузка необходимых библиотек,
- Загрузка данных в программу,
- Получение уникального списка id компаний по региону,
- Реализация функции для определения значения параметра,
- Реализация функции для расчета суммы параметров компании,
- Реализация функции для расчета темпов прироста,
- Реализация функции для нахождения ошибок в других показателях отчетности компании,
- Реализация цикла для расчета темпов прироста и проверки прочих показателей отчетности.

В результате работы программы получено однозначное соответствие компании и параметра, который определяет степень достоверности данных:

0 – все отлично, ошибок нет,
1 – темп прироста больше 100,
2 – темп прироста больше 1000,
3 – ошибка в показателях отчетности,
13 – ошибка в темпах прироста и показателях отчетности,
23 – ошибка в темпах прироста и показателях отчетности.

Данные значения параметров не являются конечным выводом, это лишь маркером для экспертов, которые будут осуществлять дальнейшую проверку показателей компании.

Финальным этапом реализации рассмотренного алгоритма является его добавление в базу данных компаний. Это позволяет в процессе анализа или пропустить компании с недостоверными данными или их индивидуально обработать. В целом, доля компаний с ошибками баланса составляет 4.46%, а с всплесками 0.52%. Не смотря на небольшой процент компаний с всплесками, 1000 кратное завышение показателей создает сильное смешение аналитических результатов.

Но работа с большими данными не останавливается только на поиске ошибок, а заключается в анализе и прогнозировании, поэтому на следующем этапе был сформирован Модуль по обработке, выявлению взаимосвязей и прогнозированию экономических панельных данных.

Методика проведения исследования включала несколько этапов. В начале была разработана система подготовки данных для применения алгоритмов машинного обучения. В рамках этого этапа проводился разведывательный анализ данных, обработка пропущенных значений и проведение тестов на стационарность и гетероскедастичность. Для дополнительной проверки необходимости логарифмирования показателей был разработан алгоритм, реализованный в рамках комплексной экспертной тестово-аналитической системы.

Следующим шагом был поэтапный процесс поиска взаимосвязей в данных. Были построены матрицы коинтегрированных показателей, а также матрица связанных показателей на основе теста Гренджера, в которой определена статистически значимая связь между соответствующими показателями, где один показатель является причиной, а другой – следствием. Затем было проведено моделирование и прогнозирование показателей с использованием алгоритмов машинного обучения и нейронных сетей.

В модельный пул были включены следующие модели машинного обучения и нейронных сетей:

- Рекуррентная нейронная сеть (RNN),
- Рекуррентная нейронная сеть с долгой краткосрочной памятью (LSTM),
- Градиентный бустинг (Gradient boosting).

В результате была разработана система автоматического поиска взаимосвязей и прогнозирования, которая была реализована в виде программного продукта. Она обладает высокой степенью автоматизации и позволяет более эффективно находить взаимосвязи между различными переменными и предсказывать будущие прогнозные значения на основе имеющихся данных.

Работа с большими данными требует не только системы их прогнозирования, но и модульных блоков их агрегирования и анализа. Поэтому в рамках развития БЗ, были разработаны аналитические блоки данных об инвестиционных проектах, компаниях и таможенной статистике, где после выбора отраслей, регионов и интересующих вас показателей пользователю предоставляются аналитические материалы в виде графиков и таблиц. Привязка всех показателей к ГИС-координатам позволяет развивать блоки ГИС-моделирования и генерировать аналитическую информацию по конкретной территории, выделенной пользователем на карте.

Все рассматриваемые модули направлены на преодоление проблем анализа больших данных и являются вспомогательными инструментами исследователя.

***

Терентьев Н.Е.

О ДОЛГОСРОЧНЫХ ДИСБАЛАНСАХ УСТОЙЧИВОГО СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ РОССИЙСКИХ РЕГИОНОВ[1]

Задачи активизации динамики экономического роста, повышения качества жизни населения, последовательной реализации долгосрочных целей национального развития России [1-2] являются особенно актуальными в рамках проводимой в настоящее время экономической политики по выстраиванию новой модели социально-экономического развития России в условиях текущих глобальных геостратегических и геоэкономических изменений. При этом крайне важным является

[1] *Работа выполнена по плану НИР ИНП РАН.*