МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ РЕАЛИЗАЦИИ АНСАМБЛЕВЫХ МОДЕЛЕЙ В ЗАДАЧЕ АНАЛИЗА СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Рассказкина Никиты Дмитриевича

Научный руководитель
доцент, к. э. н. _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2021

ВВЕДЕНИЕ

При решении задачи анализа социально-экономических показателей применяются современные методы машинного обучения, на основе которых создаются новые инструменты для аналитиков. Одним из актуальных подходов машинного обучения является использование ансамблевых методов. Методы данного типа объединяют несколько моделей для создания более мощной модели.

Особенность решаемой в данной дипломной работе задачи заключается в обработке панельных данных, для которых существует класс специальных моделей. Панельные данные образуются при наблюдении за несколькими объектами во времени и часто встречаются при анализе социально-экономических показателей. В настоящее время панельные модели активно развиваются дополняя классические методы современными подходами машинного обучения.

Целью выпускной квалификационной работы является разработка сетевого приложения, позволяющего реализовать ансамблевые модели для анализа показателей социально-экономического развития.

Задачами выпускной квалификационной работы являются:

— анализ особенностей применения ансамблевых моделей;
— получение и анализ выборки панельных данных средствами R;
— выбор, построение и настройка ансамблевых моделей;
— проектирование архитектуры и разработка приложения, реализующего прогнозирование показателей социально-экономического развития.

Объектом дипломной работы являются инструментальные средства для реализации прогностических моделей. Предметом является разработка приложения для анализа социально-экономических показателей регионального развития.

Дипломная работа состоит из введения, 3-х разделов, заключения, списка использованных источников и 10-и приложений. Общий объем работы – 67 страниц, из них 52 страниц – основное содержание, включая 11 рисунков, 12 таблиц, список использованных источников из 21 наименования.

1 Основная часть

В первом разделе были описаны понятие и виды панельных данных, осуществлён анализ моделей панельной регрессии, выявлены основные проблемы применения панельной регрессии. В данном разделе были рассмотрены виды ансамблевых моделей, представлен алгоритм на основе деревьев решений GPBoost.

Панельные данные – это данные, содержащие наблюдения о различных поперечных сечениях во времени. Объектами наблюдений могут быть страны, фирмы, отдельные лица или демографические группы. Таким образом панельные данные представляют собой набор наблюдений за одними и теми же объектами, которые производились в некоторые интервалы времени [1].

Выделяют ряд преимуществ панельных данных над простыми сечениями и временными рядами. Панельные данные могут учитывать индивидуальную неоднородность, то есть позволяют избежать ошибки спецификации. Панельные данные содержат большее количество наблюдений, что позволяет минимизировать эффект мультиколлинеарности параметров. К другим преимуществам панельных данных относят [2]:

— возможность моделировать как поведение группы объектов в целом, так и индивидуальное поведение каждого объекта;
— возможность обнаруживать и измерять статистические эффекты, недоступные для временных рядов или данных поперечного сечения;
— минимизация погрешности оценки, возникающие в результате объединения групп в один временной ряд.

При анализе панельных данных рассматриваются две основные проблемы: гетерогенное смещение и смещение самоотбора [3]. Таким образом панельные данные возможно применить для решения конкретной прикладной задачи, а именно, анализа региональной статистики, так как данную выборку возможно привести к виду панельных данных, а также повысить точность модели в сравнении с классическими методами обработки данных.

Рассмотрены основные виды моделей панельной регрессии. Среди моделей анализа панельных данных можно выделить четыре основных типа [4]:

— модель на основе метода наименьших квадратов;
— модель с фиксированным эффектом;
— модель с случайным эффектом;

— модель с смешенным эффектом.

Преимуществом модели фиксированных эффектов является то, что ошибоки могут быть коррелированы с отдельными эффектами. Если групповые эффекты не коррелируют с групповыми регрессорами, вероятно, было бы лучше использовать более простую параметризацию панельной модели.

Модель с случайными эффектами используется в тех случаях, когда объекты исследования рассматриваются как подмножество из общей выборки [5]. Преимущество случайных эффектов заключается в том, что можно включать переменные, инвариантные по времени, например, пол человека.

В данной работе будет реализована ансамблевая модель для анализа панельных данных. Рассмотрим общие принципы работы ансамблевых моделей, а также основанный на них алгоритм GPBoost. В ансамблевой модели машинного обучение вводится понятие базовых моделей. Базовые модели – это модели, используемые для проектирования более сложных моделей путём объединения нескольких из них. Как правило базовые модели имеют либо высокое смещение, либо большой разброс. В таких случаях идея ансамблевых методов состоит в том, чтобы уменьшить смещение и/или разброс таких базовых моделей, объединяя их в одну ансамблевую модель, которая достигает лучших результатов. Выделяют три основных типа метаалгоритмов, которые направлены на объединение базовых методов [6]:

— бэггинг – параллельное и независимое обучение однородных базовых моделей с последующим объединением посредством некоторого детерминированного процесса усреднения;
— бустинг – последовательное обучение (базовая модель зависит от результатов обучения предыдущих) однородных базовых моделей с последующим объединением посредством некоторого детерминированного процесса усреднения;
— стекинг – параллельное обучение разнородных базовых моделей с последующим обучением ансамблевой модели для вывода прогноза на основании предсказаний включенных в неё базовых моделей.

В данной работе будут использоваться только однородные модели, что означает что будет рассматриваться только бэггинг и бустинг. В качестве модели для осуществления регрессии взят ансамбль решающих деревьев [7]. Помимо высокой точности прогнозирования, бустинг деревьев обладает сле-

дующими преимуществами:

- возможность моделирование нелинейностей, разрывов и сложных взаимодействий высокого порядка;
- устойчивость к выбросам и мультиколлинеарности между переменными-предикторами;
- масштабная инвариантность к монотонным преобразованиям переменных-предикторов;
- возможность обработки пропущенных значений в переменных предиктора.

Для моделей на основе панельных регрессий предлагается использовать алгоритм GPBoost, где функция регрессии $F(X)$ строится с использованием ансамбля деревьев. Гибридизация методов, как правило, даёт более точные модели, частности объединение деревьев решений в единую ансамблевую модель. Метод GPBoost в дальнейшем будет исследован на возможность применения для анализа социально-экономических данных, имеющих панельную структуру.

При оценке моделей в данной выпускной квалификационной работе будут использоваться метрики SMAPE и RMSE. Выбор SMAPE метрики обусловлен её простой интерпретацией, а также устойчивостью к выбросам и величине фактического значения. RMSE – классический метод оценки точности регрессии, по которому возможно выявить смещение SMAPE в случае его появления [8].

Второй раздел посвящен разработке приложения. Был обоснован выбор языка R для анализа панельных данных, описана реализация различных инструментов анализа панельных данных. Предложена архитектура приложения с использованием клиент-серверной технологии.

R – язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда для вычислений с открытым исходным кодом в рамках проекта GNU. При реализации будет использоваться версия языка R 3.6 [9]. В качестве среды разработки была выбрана свободно распространяемая RStudio версии 1.4. Помимо компиляции и синтаксического анализа кода данная среда разработки обладает рядом возможностей, упрощающих и ускоряющих процесс разработки приложения.

Наиболее популярным и хорошо задокументированным пакетом язы-

ка R для обработки панельных данных является пакет plm. Данный пакет позволяет осуществлять предобработку панельных данных, включает набор базовых методов обработки панельных данных и тестов для оценки моделей.

Алгоритм GPBoost не входит в данных пакет. Для использования GPBoost необходимо установить дополнительный одноименный пакет. GPBoost – компилируемый пакет, написанный на C++, что позволяет увеличить скорость вычислений. Наиболее значимыми параметрами являются:

— количество решающих деревьев ансамбля;
— шаг обучения;
— максимальная глубина дерева;
— минимальное количество объектов в листе решающего дерева.

Для визуализации данных использовались пакеты car и gplots. Инструментами пакета car возможно построить график значений атрибутов с добавлением аппроксимации между точками для большей наглядности. Для разработки приложения для анализа социально-экономических показателей ансамблевыми моделями была выбрана клиент-серверная архитектура, так как она обеспечивает ряд преимуществ по сравнению с автономным приложение, при этом приложение не подразумевает необходимость автономной работы, а пропускной способности стандартных каналов связи хватает для решения поставленной задачи.

Было принято решение реализовать серверную часть на языке – Python версии 3.9 [10]. Для реализации серверной части приложения был выбран микро-фреймворк Flask. Микро в данном случае означает модульность фреймворка. Изначально Flask предоставляет только базовые серверные возможности, однако при необходимости их можно расширить дополнительными пакетами. Так для более корректной реализации REST стандарта обращения к API установлено дополнение Flask-RESTful.

Поскольку серверная часть написана на двух языках программирования, необходимо определить их способ взаимодействия. Данная библиотека обладает всем функционалом, необходимым для реализации поставленных задач, при этом её производительность является достаточной для корректной работы модуля, содержащего R код.

Для развёртывания серверного модуля используется среда виртуализации Docker. Docker – программное обеспечение для автоматизации развёрты-

вания и управления приложениями в средах с поддержкой контейнеризации. Позволяет создать виртуальную среду для программного решения, включающую его окружение и зависимости, которая помещается в контейнер.

Модуль пользовательского приложения реализован в отдельном репозитории. Данное приложение представляет собой Windows-приложение. Для реализации используется Python с установленным фреймворком PyQT. PyQt — набор расширений графического фреймворка Qt для языка программирования Python, выполненный в виде расширения Python. PyQt практически полностью реализует возможности Qt. Для создания пользовательского интерфейса использовалась среда Qt Designer. Данная среда позволяет с помощью графического интерфейса создавать пользовательский интерфейс приложения.

Таким образом получившееся программный комплекс имеет приложение с пользовательским интерфейсом, которой отправляет HTTP запросы к REST API серверу и визуализирует полученные ответы. Отправленные запросы принимает контейнер Docker, и переадресует его на указанный порт операционной системы ubuntu. Затем запрос обрабатывается gunicorn-сервером и передаётся фреймворку Flask, который передаёт информацию из запроса обработчику. Если информация обрабатывается средствами R, тогда предварительно она преобразуется средствами rpy2 в формат данных R и обратно. Преобразованные данные отправляются на клиентское приложение и визуализируются. Полученная архитектура имеет вид как на рисуноке.

В процессе проектирования приложения для реализации прогнозных моделей предложено использовать фреймворк Flask в качестве сервера, фреймворк PyQt5 для реализации пользовательского интерфейса и Docker для развёртывания сервера. Подобный подход обеспечит взаимодействие интерфейсной части и пакетов R.

Перед началом разработки сервера необходимо определить набор методов, посредством которых будет осуществляться обращение к серверу. Для REST API сервера такими методами являются URL-маршруты, каждый из которых однозначно определяет запрашиваемый ресурс.

После создания проекта PyCharm необходимо определить базовую структуру проекта, а именно расположение и вложенность всех модулей сервера. После создания проекта пустого проекта, репозиторий будет содержать одну

папку venv. Данная директория хранит созданное автоматически виртуальное окружение Python. В корне проекта были созданы следующие директории и файлы:

— `app.py` – исполняемый файл сервера;
— `requirements.txt` – файл хранящий все зависимости интерпретатора виртуального окружения;
— `responses.py` – модуль, хранящий шаблоны ответов сервера;
— `exceptions.py` – модуль ошибок обрабатываемых сервером;
— `validation` – директория отведённая для модуля валидации запросов;
— `tests` – директория модуля автоматического тестирования сервера;
— `services` – директория модуля сервисов сервера.

Все объекты, с которыми может взаимодействовать пользователь написаны на языке R, поэтому они были вынесены в отдельный сервис R, который содержит реализации данных объектов.

Для разработки пользовательского интерфейса был создан отдельный репозиторий. Программная разработка пользовательского модуля осуществляется в среде разработки PyCharm, а конструирование графических элементов происходит производилось в среде Qt Disigner.

Первый этап разработки заключается в проектировании логики взаимодействия пользователя с приложением и последующего создания формы приложения, которое реализует вышеуказанную логику. Сохранённая форма является файлом с расширением .ui, который содержит yaml-структуру, описывающую все элементы формы. Была создана форма form.ui, размещённая в корне проекта пользовательского модуля.

Затем был создан файл textttmain.py. Данный файл открывает окно формы при запуске и назначает каждому элементу функции обработки событий, реализующие их функционал. Функции обработки событий были вынесены в отдельный файл. Для отправки запросов на сервер используется библиотека requests.

Таким образом была произведена разработка клиент-серверного приложения, позволяющего реализовать добавление новых методов прогнозирования, обеспечивающего возможность использования приложения на разных пользовательских платформах, интеграцию разработанных модулей в существующую информационно-аналитическую систему.

В третьем разделе представлено разработанное приложение, позволяющее применить ряд моделей прогнозирования, в том числе GPBoost. Реализован вычислительный эксперимент по выбору прогностической модели. Рассмотрена прикладная задача прогнозирования инновационного развития отдельных регионов Российской Федерации.

Будет проведён вычислительный эксперимент, в котором будут сравнены различные модели панельных данных. Для проведения эксперимента используется выборка, полученная из информационно-аналитической системы FIRA [11]. В качестве прогнозируемого показателя был выбран объём инновационных товаров в регионе. Данный выбор обусловлен тем, что данный показатель отражает результативность инновационного развития региона. Для обучающей выборки были выбраны 7 показателей: соотношение заемного и собственного капитала, обеспеченность собственными оборотными средствами, текущая ликвидность, доля кредитов и займов в краткосрочных пассивах, доля долгосрочных обязательств в совокупном капитале, фондоотдача, оборачиваемость активов.

Предлагается следующая методика вычислительного эксперимента. Описанный ранее ансамблевый алгоритм GPBoost для исследования панельных данных будет применён для ряда моделей. Данные модели будут отличаться значениями стандартных параметров алгоритма, а именно:

— p_1 – количество решающих деревьев ансамбля;
— p_2 – максимальная глубина дерева;
— p_3 – минимальное количество объектов в листе решающего дерева.

Построенные модели будут сравниваться по метрикам RMSE и SMAPE, а также по времени, которое было затрачено на обучение модели. Было проведено сравнение полученной GPBoost модели с классическими методами регрессии панельных данных. По его итогам модель GPBoost показывает наилучшую точность предсказания.

Полученная модель может быть использована для прогнозирования как объёма инновационных товаров, так и других показателей регионов. Предложенный подход на основе гибридных методов панельной регрессии и деревьев решений обеспечил достаточно высокую точность прогноза относительно классических панельных методов.

Функционал разработанного приложения предполагает выполнение сле-

дующих этапов анализа данных:

— ввод данных в формате csv;
— реализация запросов на выборку данных по периодам;
— применение статистических тестов для оценки гетероскедастичности индивидуальных и временных эффектов;
— выбор прогнозного алгоритма;
— настройка параметров GPBoost.

Интерфейс разработанного приложения имеет три вкладки: «Данные», «Анализ», «Регрессия».

Представленное приложение обеспечивает удобный способ реализации ансамблевого подхода в задаче прогнозирования для панельных данных. Расширенный функционал позволяет реализовать основные этапы моделирования с учетом проверки гетероскедастичности и вариабельности атрибутов выборки. К преимуществам данного приложения можно отнести обеспечение возможности сравнения различных панельных моделей.

ЗАКЛЮЧЕНИЕ

В данной дипломной работе был проведён анализ методов обработки панельных данных. Преимущества панельных методов заключается в возможности учёта смещения, вызванного не вошедшими в выборку факторами или неполнотой выборки. Помимо этого панельные модели показывают хорошие результаты при относительно небольшом объёме выборки, что характерно для задач прогнозирования, связанных с региональной статистикой.

Были рассмотрены модели панельных данных с фиксированными и случайными эффектами, а также гибридная ансамблевая модель GPBoost. Модель GPBoost позволяет объединить методы бустинга на основе деревьев решений и панельную регрессию.

Было разработано приложение, обеспечивающее как реализацию обычных методов панельной регрессии, так и модификацию метода с помощью ансамбля деревьев решений.

Архитектура приложения предполагает использование отдельных пакетов R для анализа данных. Серверная часть приложения реализована на платформе Python. Пользовательский интерфейс реализован с помощью фреймворка PyQt5.

Расширенный функционал приложения обеспечивает загрузку и отбор данных, выполнение статистических тестов на гетероскедастичность, выбор и настройку панельных методов. Особенностью данного приложения является возможность сравнения различных моделей по отдельным показателям ошибки.

Апробация приложения осуществлялась для набора показателей, характеризующих экономическое развитие регионов Российской Федерации. Прогнозирование объёма инновационных товаров как важнейшего показателя регионального развития выполнялась с помощью ряда моделей. С точки зрения точности моделей наиболее предпочтительной оказалась модель, построенная на основе алгоритма GPBoost.

Вычислительный эксперимент подтвердил целесообразность применения ансамблевых моделей для данной прикладной задачи. Направлением дальнейших исследований является использование в задачах прогнозирования различных наборов социально-экономических показателей регионального развития.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Айвазян, С. А. Методы эконометрики / С. А. Айвазян. -М: ИНФРА-М, 2019. - 235 с.

2 Greene, W. H. Econometric Analysis / W. H. Greene. - Upper Saddle River: Prentice Hall, 2003. - 304 p.

3 Ратникова, Т.А. Введение в экономический анализ панельных данных / Т. А. Ратникова // Экономический журнал ВШЭ, 2006. - Т. 10, № 2. -С. 274-276.

4 Matthew, G., George, K., Lloyd, E. Fixed and random effects models / G. Matthew, K. George, E. Lloyd // Wiley Interdisciplinary Reviews: Computational Statistics, 2011. - T. 3, №4, - Pp. 181–190.

5 Nakale S., Kleyn, J., Arashi, M. Feasible generalised least squares estimators in serially correlated error models from an asymmetry viewpoint / S. Nakale, J. Kleyn, M. Arashi. // Peer-reviewed Proceedings of SASA, 2013. -Pp. 54-56.

6 Susan, A., Mohsen, B., Guido, I., Zhaonan Q. Ensemble Methods for Causal Effects in Panel Data Settings / A. Susan, B. Mohsen, I. Guido, Q. Zhaonan // AEA Papers and Proceedings, 2019. - T. 109, - Pp. 65-70.

7 Sigrist, F. Gaussian Process Boosting / F. Sigrist // arXiv preprint arXiv:2004.02653, 2020.

8 Чернышова, Г. Ю., Самаркина, Е. А. Методы интеллектуального анализа данных для прогнозирования финансовых временных рядов / Г. Ю. Чернышова, Е. А. Самаркина // Изв. Сарат. ун-та. Нов. сер. Сер. Экономика. Управление. Право - 2019. -Т. 19, №2. - С. 181-188.

9 Download R: [Электронный ресурс] URL: https://cran.r-project.org/bin/windows/base/ (дата обращения 09.12.2020)

10 Download Python: [Электронный ресурс] URL: https://www.python.org/downloads/ (дата обращения 12.03.2020)

11 FIRA: [Электронный ресурс] URL: https://fira.ru/ (дата обращения 10.20.2020)