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 МЕТОДОВ DATA MINING В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Черникова Виктора Алексеевича

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

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

Саратов 2021

ВВЕДЕНИЕ

Задачи прогнозирования и построения предикативных моделей остаются актуальными на сегодняшний день, и для их решения существует широкий спектр методов, как статистических, так и методов интеллектуального анализа данных и машинного обучения. С учетом разнообразия решаемых задач подбор конкретных методов является важным этапом. Кроме того, активно развиваются подходы основанные на современных алгоритмах, использующих гибридные методы. В частности, в условиях недостаточности исходных данных широко применяются специфические алгоритмы, в частности панельные регрессии. Преимущество панельной регрессии в отличие от остальных методов заключается в работе с ограниченными по объему выборками, что делает их более применимыми к социально - экономическим показателям представленным в рамках региональной статистики. Актуальность этих моделей обусловлена независимостью от объема исходных данных.

Целью данной выпускной квалификационной работы является разработка приложения для выбора и применения методов прогнозирования социально - экономических показателей. Для достижения заданной цели требуется выполнить следующие задачи:

— классифицировать методы прогнозирования;
— осуществить анализ инструментальных средств для реализации прогностических моделей;
— сформировать выборку и осуществить препроцессинговую обработку данных;
— разработать приложение для прогнозирования регионального развития;
— апробировать приложение для показателей экономического развития регионов.

Предметом данной бакалаврской работы является применение прогностических методов в рамках задачи экономико - математического моделирования. Объектом данной бакалаврской работы являются инструментальные средства для реализации методов прогнозирования.

Бакалаврская работа состоит из введения, трех разделов, заключения, списка использованных источников и десяти приложений. Общий объем работы – 68 страниц, из них 40 страниц – основное содержание, включая 14 рисунков и список использованных источников из 21 наименования.

1 Основная часть

Первый раздел бакалаврской работы посвящен классификации и описанию примянемых методов прогнозирования. Данный раздел содержит три подраздела.

Первый подраздел содержит сравнительный анализ моделей панельной регрессии и других количественных методов прогнозирования. Все современные методы прогнозирования относятся к одному из двух видов:

— количественные методы прогнозирования - построенные на математических измерительных моделях;

— качественные методы прогнозирования - основанные на человеческих суждениях, имеющие субъективный характер.

Рассматриваются следующие методы прогнозирования:

— стохастические модели Брауна и Хольта - Винтерса – учитывают стохастическую природу исходных данных, а также аддитивную и мультипликативную сезонность;

— методы панельной регрессии [1–3], такие как модель сквозной регрессии, модель регрессии с детерминированным индивидуальным эффектом и модель регрессии со случайным индивидуальным эффектом – учитывают панельную структуру исходных данных;

— авторегрессионные модели [4, 5] ARMA и ARIMA – являются наиболее простыми для построения моделями.

Ранние эконометрические модели, опирающиеся на данные пространственных выборок или временных рядов, носили агрегированный характер и описывали поведение усредненных объектов. Со временем выяснилось, что эти модели часто оказывались не слишком эффективными инструментами для анализа экономических явлений и выработки рекомендаций по социально-экономической политике. Очень часто ни значения, ни знакие коэффициентов, посчитанных по регрессиям для агрегированных временных рядов, не соответствовали предположениям экономической теории, так как возникало серьезное смещение агрегирования. Появление новых моделей экономических явлений(моделей анализа панельных данных) обеспечивают разнообразные возможности учета неоднородности. Данные панельного типа предоставляют исследователю большое количество наблюдений, увеличивая число степеней свободы и снижая зависимость между объясняющими переменными, а следо-

вательно, стандартные ошибки оценок. Появляется возможность анализировать множество экономических вопросов, которые не могут быть адресованы к временным рядам и пространственных данным в отдельности. Позволяют предотвратить смещение агрегированности, неизбежно возникающее как при анализе временных рядов (где рассматривается временная эволюция усредненного репрезентативного объекта), так и при анализе перекрестных данных (где не учитываются ненаблюдаемые индивидуальные характеристики объектов и предполагается однородность, всех коэффициентов регрессии). С их помощью можно проследить индивидуальную эволюцию характеристик всех объектов выборки во времени. Они решают проблему поиска пригодных инструментов при оценивании моделей с эндогенными (т.е. коррелированными со случайными ошибками) регрессорами.

Второй подраздел содержит общую информацию о методах Data Mining, которые применяются в работе. Одним из самых известных методов машинного обучения [6] являются деревья решений, позволяющие решать задачи классификации и регрессии. Популярность данного метода обусловлена простотой его устройства и гибкостью применения, но он чувствителен к исходным данным, на которых обучается модель. Также деревья решений вычислительно дорого обучать, так как присутствует большой риск переобучения модели. Чтобы устранить все риски и недостатки, используется комитет случайных деревьев [7]. Деревья комитета работают параллельно и не взаимодействуют между собой. Данный метод строит множество деревьев решений во время обучения и выводит среднее прогнозное значение от всех индивидуальных деревьев. Результат работы комитета случайных деревьев является метаоценкой, так как комбинирует множество прогнозов, что приводит к некоторым модификациям.

1. Число характеристик, которое может быть разделено по каждому дереву ограничено некоторым соотношением к общему числу характеристик. Это позволяет не полагаться слишком сильно на конкретные характеристики и в равной степени использовать всю выборку.
2. Каждое дерево берет случайные части выборки, что предотвращает переобучаемость.

Аналогично комитету случайных деревьев, метод k-ближайшего соседа применяется в задачах классификации и регрессии. Он использует схожесть

характеристик для прогнозирования значений. Это означает, что значение следующей точки вычисляется на основе ближайших точек обучающей выборки. Ключевым моментом в работе метода KNN является параметр k, от которого зависит итоговый результат. В большинстве случаев k выбирается эмпирически, путем сравнения величин ошибок.

Третий подраздел содержит описание способов оценки регрессионных моделей, таких как:

— средняя квадратическая ошибка MSE – самый простой и распространенный показатель для оценки регрессии, но наименее полезный;
— средняя абсолютная ошибка MAE – менее чувствительна к большим ошибкам по сравнению с MSE;
— коэффициент детерминации R^2 – является более точной оценкой регрессионной модели, чем MSE и MAE, так как дает не среднее значение ошибки, а коэффициент;
— скорректированный R^2 – исправляет проблему возрастания коэффициента с увеличением количества независимых переменных;
— средние квадратическая и абсолютная ошибки, выраженные в процентах MSPE и MAPE – дают не абсолютные значения ошибок, а относительные;
— среднеквадратическая логарифмическая ошибка RMSLE – используется в тех же ситуациях, что и MSPE и MAPE, то есть для получения относительных значений ошибки.

Во втором разделе проводится анализ инструментальных средств для разработки приложения. Данный раздел состоит из двух подразделов.

Первый подраздел содержит описание языка программирования Python, а также обзор необходимых библиотек для разработки приложения. Рассматриваются следующие библиотеки:

— pandas [8] – библиотека для работы с табличными данными;
— linearmodels [9] – библиотека, в которой реализованы модели панельной регрессии;
— scikit-learn [10] – библиотека, в которой реализованы методы Data Mining;
— statsmodels – библиотека для проведения стандартных статистических тестов;
— matplotlib – библиотека для визуализации данных с помощью двумер-

ной графики.

Во втором подразделе описывается архитектура разрабатываемого приложения. Приложение состоит из следующих модулей:

— модуль очистки данных – проводит удаление ненужных строк в таблицах;
— модуль обработки пропущенных значений – заполняет пропуски во временных рядах;
— модуль нормализации данных – нормализирует данные к унифицированному виду;
— модуль применения стандартных статистических тестов – проводит необходимые тесты для выбора наиболее подходящей модели панельной регрессии;
— модуль реализации методов панельной регрессии – применяет методы панельной регрессии на обработанных данных;
— модуль реализации методов Data Mining – применяет методы Data Mining на обработанных данных.

Интерфейс приложения представляет собой окно с функциональными кнопками и рабочей областью. Кнопки позволяют выбрать файл для вывода значений и очистить рабочую область. Таким образом интер-фейс приложения необходим для асинхронного мониторинга результатов работы модулей приложения.

Третий раздел посвящен проведению вычислительного эксперимента. Данный раздел состоит из трех подразделов.

Первый подраздел содержит описание выборки и предобработки данных. Исходные данные были взяты из информационно-аналитической системы FIRA(Первое Независимое Рейтинговое Агенство) и открытого источника Росстата. FIRA работает сфере информационно-аналитической поддержки российских и зарубежных компаний. Данные представляют собой набор таких показателей как индекс физического объема, импорт, экспорт, прибыль от продаж и фондоотдача. Данные из разных источников имеют отличающийся формат и длину временного ряда, следовательно вся исходная выборка требует форматирования, корректировки и нормализации. Для заполнения пропущенных значений реализован метод взвешенного среднего. Для нормализации данных применено десятичное масштабирование.

Второй подраздел содержит описание и применение стандартных статистических тестов, а также анализ их результатов. Были применены следующие тесты:

— графический тест на гомоскедастичность;
— тест Уайта на гомоскедастичность;
— тест Бройша - Пагана на гомоскедастичность;
— тест Дарбина - Уотсона для определения присутствия автокорреляции.

Для графического метода необходимо загрузить в самую простую модель панельной регрессии независимую и зависимую переменную и построить прогноз. По оси абсцисс выводятся спрогнозированные значения, по оси ординат – остаточные ошибки. Таким образом можно проследить однородность данных, загруженных в модель. Для уточнения выводов следует воспользоваться тестами Уайта, Бройша - Пагана и Дарбина - Уотсона. Тесты Уайта и Бройша - Пагана - это универсальная процедуры тестирования гетероскедастичности случайных ошибок линейной регрессионной модели. В результате теста Дарбина - Уотсона получается статистический критерий, используемый для тестирования автокорреляции первого порядка элементов исследуемой последовательности. В результате применения тестов Уайта и Бройша - Пагана, уровень достоверности p превышает 0,05 только у индекса физического объема и фондоотдачи, в остальных случаях гомоскедастичность отсутствует. В результате применения статистических тестов данные по индексу физического объема и фондооотдаче имеют гомоскедастичный или однородный характер, все остальные данные гетероскедастичны. Также после применения теста Дарбина - Уотсона оказалось, что вся выборка имеет положительную автокорреляцию. Таким образом, для сформированной выборки предлагается использовать модели с фиксированными и случайными эффектами.

Третий подраздел содержит построение прогнозной модели на примере отдельных показателей развития регионов. Была реализована следующая схема вычислительного эксперимента:

1. применение моделей панельной регрессии для всех независимых переменных;
2. применение моделей панельной регрессии для каждой независимой переменной;
3. использование методов Data Mining Random Forest и k-Nearest Neighbour

для всех независимых переменных;

4. оценка моделей панельной регрессии, сравнение с методами Data Mining Random Forest и k-Nearest Neighbour.

Чтобы выбрать конфигурацию моделей Data Mining необходимо сравнить среднеквадратические ошибки, полученные в результате изменения этой конфигурации, и выбрать конфигурацию с наименьшим значением ошибки. Таким образом были выбраны значения:

— 3 дерева для метода Random Forest;

— 3 соседа для метода k-Nearest Neighbour.

Введем обозначения моделей для составления таблицы с результатами эксперимента:

— M_1 - M_3 – модели со случайными и временными эффектами, а также объединенная модель, в которые загружены все независимые переменные;

— M_4 - M_{18} – модели со случайными и временными эффектами, а также объединенные модели, в которых используется только одна независимая переменная;

— DM_1, DM_2 – модели с применением методов Data Mining Random Forest (DM_1) и k-Nearest Neighbour (DM_2).

Полученный результат включает в себя вычисленные коэффициенты регрессии для различных моделей, уровень достоверности p, коэффициент детерминации R^2. Модели $M_1 - M_3$ имеют достаточно высокий коэффициент детерминации $R^2 > 0,9$, то есть хорошо объясняют полученную модель данных. При этом уровень достоверности $p < 0,05$.

M_1 и M_2 практически не отличаются по оценкам R^2 и уровню достоверности p, но единственная проблема - интерпретация коэффициентов регрессии. Объединенная модель также имеет большой R^2, но у индекса физического объема и фондоотдачи $p > 0,05$. Это означает, что объединенная модель не подходит для такой выборки, что было показано на этапе тестирования. Также это подтвердилось в объединенных моделях с отдельными независимыми переменными. Модели, у которых $R^2 < 0,6$ не рассматриваются.

Анализ построенных регрессионных моделей свидетельствует о том, что валовый региональный продукт больше всего зависит от прибыли от продаж. Но на самом деле ВРП может зависеть от неучтенных показателей даже в

большей степени, чем от прибыли от продаж. Следовательно, необходимо дальнейшее исследование.

Методы панельной регрессии по точности прогноза сравнимы с методами Data Mining, но они позволяют увидеть более подробную оценку влияния независимых переменных в модели.

ЗАКЛЮЧЕНИЕ

В рамках выпускной квалификационной работы для прогнозирования по-казателей регионального развития предложено использовать методы панельной регрессии. Архитектура реализованного приложения предполагает использование отдельных модулей на платформе Python.

В качестве выборки используются данные из открытых источников, включающие показатели регионального развития за 2002 - 2019 гг. Разработанное приложение предоставляет следующий функционал:

— загрузить выборку;
— выполнить предварительную обработку данных, включающую в себя удаление объектов, для которых в открытых источниках отсутствуют необходимые значения показателей, заполнение пропущенных значений и нормализацию;
— применить различные методы панельной регрессии, такие как объединенная модель панельных данных и модели с фиксированными и случайными эффектами;
— осуществить вывод результатов, включающий оценки моделей, такие как коэффициенты регрессии, коэффициент детерминации и уровень достоверности.

Традиционные корреляционные модели не позволяют оценить связи между факторами в задаче анализа макроэкономических систем, таких как регионы. И поэтому следует предпринять попытки использования расширенного эконометрического инструментария для анализа взаимосвязей факторов при оценке показателей регионов.

В результате были получены прогнозные модели для оценки валового регионального продукта, которые свидетельствуют о том, что такие показатели регионального развития как индекс физического объема и фондоотдача влияют на него в гораздо меньшей степени, чем импорт, экспорт и прибыль от продаж. Но на самом деле ВРП может зависеть от других показателей регионального развития в большей степени, чем от содержащихся в выборке. Это значит, что для получения полной картины необходимо провести больше экспериментов с другими конфигурациями выборки.

Полученные значения коэффициентов регрессии могут свидетельствовать о том, что необходимо изменить принцип подбора факторов. Дальнейшее

исследование будет связано с переконфигурированием выборки, расширением списка факторов и пошагового анализа различных наборов независимых переменных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Айвазян, С.А. Методы эконометрики: учебник / С.А. Айвазян. – М. Магистр: ИНФРА-М, 2010. –512 с.

2 Айвазян, С.А. Прикладная статистика: классификация и снижения размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.

3 Елисеева, И.И. Эконометрика: учебник для магистров / И.И. Елисеева. – М. : Издательство Юрайт, 2014. – 453 с. – Серия : Магистр.

4 Norizan, M. Short Term Load Forecasting Using Double Seasonal ARIMA Model / M. Norizan, A. Maizah Hura, I. Zuhaimy // Regional Conference on Statistical Sciences, Malaysia, Kelantan, 2010. Pp. 57 - 73.

5 Conejo, A.J. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models / A.J. Conejo // IEEE transaction on power systems. 2005, Vol. 20, No. 2. Pp. 1035 - 1042.

6 Hastie, T. Chapter 15. Random Forests / T. Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning: Data Mining, Inference, and Prediction – 2nd ed. – Springer-Verlag, 2009. – 746 p.

7 Chakure, A. [Электронный ресурс] Medium : Random Forest Regression / A. Chakure. URL: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f (дата обращения: 21.05.2021). - Загл. с экрана. - Яз. англ.

8 Документация библиотеки pandas [Электронный ресурс] URL: https://pandas.pydata.org/pandas-docs/stable/index.html (дата обращения: 20.05.2021)

9 Документация библиотеки linearmodels [Электронный ресурс] URL: https://bashtage.github.io/linearmodels/index.html (дата обращения: 20.05.2021)

10 Документация библиотеки scikit-learn [Электронный ресурс] URL: https://scikit-learn.org/stable/index.html (дата обращения: 20.05.2021)