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

АНАЛИЗ ИНДИКАТОРОВ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ РАЗВИТИЯ РЕГИОНОВ РОССИЙСКОЙ ФЕДЕРАЦИИ В ЗАДАЧЕ СНИЖЕНИЯ УРОВНЯ БЕДНОСТИ И ПОВЫШЕНИЯ БЛАГОПОЛУЧИЯ ЛЮДЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Федоровой Алёны Дмитриевны

Научный руководитель
доцент, к. ф.-м. н. _____ О. В. Мещерякова

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

Саратов 2023

ВВЕДЕНИЕ

Борьба с бедностью является одним из важнейших направлений государственной социально-экономической политики в России. В периоды глобальных экономических и политических кризисов масштабы бедности возрастают, что негативно сказывается как на социальной стабильности, так и на дальнейшем векторе общественного развития. Именно поэтому в наше время, на фоне мировых экономических потрясений проблема поиска и применения наиболее эффективных механизмов борьбы с бедностью и смягчения ее негативных факторов приобретает особую актуальность.

Основные ориентиры государственной политики по снижению уровня бедности населения обозначены в Указе Президента РФ от 21 июля 2020 года № 474 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2030 года». Одним из ключевых целевых индикаторов на ближайшее десятилетие определено двукратное снижение показателя «Уровень бедности» к 2030 году. Для достижения поставленных целей субъекты Российской Федерации разрабатывают и реализуют региональные программы по снижению доли малоимущего населения (далее – региональные программы), в которых учтены социально-экономические, географические и инфраструктурные особенности регионов. Принимая во внимание тот факт, что по всем мировым рейтингам наша страна не является бедной, к сожалению, по уровню бедности она не входит даже в первые 50 стран мира.

Целью работы является оценка индикаторов показателя уровня бедности определенного в “Едином плане по достижению национальных целей развития Российской Федерации на период до 2024 года и плановый период до 2030 года” от 21 июля 2020 года (далее именуемый как Единый план) с использованием аналитики больших данных. Среди задач, решаемых в работе, стали:

— задача сбора данных;
— выбор и использование инструментов анализа данных;
— построение модели и использование аналитики больших данных, связанных с задачами оценки индикаторов показателя уровня бедности, что позволяет прогнозировать значение показателей в различных регионах РФ и определять факторы, сильно связанные с этим показателем.

В основе работы лежат открытые данные, предоставляемые государ-

ственными источниками, такими как "Открытые данные России ИНИД (Инфраструктура научно-исследовательских данных), Росстат, Единая межведомственная информационно-статистическая система и т.п.

В качестве предмета исследования был взят такой показатель как уровень бедности. И для решения поставленных задач необходимо проведение статистического анализа, построение модели и её обучение для выявления закономерностей, связей и тенденций, связанных с оценкой индикаторов показателя бедности. Полученные оценки позволяют выделить регионы со значительным расхождением от показателей Единого плана.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой части (Понятия, подходы, методы и инструменты) объясняется понятие "бедность рассказывается о методах ее измерения и множестве показателей, влияющие на уровень бедности населения.

Бедность – это особое состояние, когда люди или семьи не обладают достаточными доходами для обеспечения необходимых для жизнедеятельности потребностей и уровня потребления, которые являются общественно признанными. Существуют различные подходы к измерению бедности.

Основными индикаторами данного показателя являются: среднемесячный доход людей в области и черта бедности измеряемые в рублях, инфляция, валовой региональный продукт (ВРП), экспорт и импорт области, измеряемый в долларах, общие расходы на здравоохранение на душу населения, продолжительность жизни населения, суммарный коэффициент рождаемости (число детей, которые родились бы у каждой женщины, если бы текущие показатели рождаемости по возрасту оставались неизменными).

Для изучения использовались технологии Big data. Big data — комплекс подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

Обработка данных - это процесс преобразования информации от исходной ее формы к виду, необходимому для использования в целях получения определенного результата. Сбор, накопление, хранение информации, зачастую не последний этап информационного процесса.

Учитывая основные характеристики Big data, для дальнейшей работы необходимо было провести обработку этих данных. Big data представляет собой огромные объемы данных, которые невозможно эффективно обработать с помощью традиционных методов и инструментов. Однако, благодаря прогрессу в области машинного обучения, стали доступны новые подходы и технологии для анализа и извлечения ценной информации из этих объемных данных. Существующая практика использования Big data продемонстрирова-

ла методы и подходы, которые используются для работы с такими объемными данными.

Во второй части выбрана аналитика больших данных в решении социальных и экономических задач. Использование больших данных становится все более распространенным в различных сферах жизни общества, и государственное управление не является исключением. Большие объемы данных предоставляют уникальную возможность для анализа и принятия обоснованных решений.

К задачам государственного управления относятся такие задачи, как:

— Анализ социально-экономических данных.
— Улучшение качества государственных услуг.
— Прогнозирование и принятие решений.

В настоящее время анализ больших данных играет ключевую роль в поддержке процесса принятия решений в государственном управлении. Он позволяет использовать объемные, разнообразные и динамичные данные для оценки эффективности и прогнозирования результатов государственных программ. Понимая важность эффективного распределения ресурсов, и стремясь достичь национальных целей, государственные учреждения могут использовать анализ больших данных для принятия обоснованных решений о продолжении финансирования государственных программ, и в конечном итоге - повысить эффективность реализации национальных стратегий и решения поставленных задач.

“Единый план по достижению национальных целей развития Российской Федерации на период до 2024 года и плановый период до 2030 года” подписанный 21 июля 2020 года, является ключевым документом, определяющим стратегические приоритеты и направления развития страны на среднесрочную и долгосрочную перспективы. Созданный с учетом мировых трендов и основных задач, стоящих перед страной, данный план является важным инструментом государственного управления и позволяет определить приоритетные секторы развития, цели и задачи национального уровня.

Основной целью Единого плана является обеспечение устойчивого и сбалансированного развития Российской Федерации, а также отдельных регионов, в соответствии с задачами, поставленными национальной стратегией. Стратегия предполагает достижения следующих национальных целей:

1. Сохранение населения, здоровье и благополучие людей.
2. Возможности для самореализации и развития талантов.
3. Комфортная и безопасная среда для жизни.
4. Достойный, эффективный труд и успешное предпринимательство.
5. Цифровая трансформация.

Этот план является основой для формирования информационной системы мониторинга достижения национальных целей развития. Такая система позволит отслеживать ход достижения целей развития, а также выявлять и анализировать причины отклонений, своевременно корректировать необходимые действия и мероприятия как на федеральном, так и на региональном уровнях.

В третьей, практической части, дипломной работы говорится об о способе оценивания благополучия регионов, построении модели, её обучении и нахождении регионов имеющие критические индикаторы показателей, ведущих население к бедности.

В практической части используется технологии Big data, программные решения выполнены средствами языка Python и его библиотек, сами данные были взяты из открытых государственных источников. В качестве изучения был сформирован набор данных, содержащий в себе информацию об основных социально-экономических индикаторов, которые влияют на уровень бедности в 86 субъектах Российской Федерации. Основными индикаторами исследования являются: среднемесячный доход людей в области и черта бедности, измеряемые в рублях, инфляция, валовой региональный продукт (ВРП), экспорт и импорт области, измеряемые в долларах, общие расходы на здравоохранение на душу населения, продолжительность жизни населения, суммарный коэффициент рождаемости (число детей, которые родились бы у каждой женщины, если бы текущие показатели рождаемости по возрасту оставались неизменными), число умерших детей (число детей, умерших в возрасте до 1 года). Датасет состоит из 87 строк и 11 столбцов, не имеющий нулевых значений. Повторяющихся значений в наборе данных нет.

Для визуализации данных используются гистограммы, а для поиска выбросов используются графики boxplot. Предварительный анализ распределения данных показал, что некоторые характеристики сильно разбросаны и поэтому они не будут удаляться. Для целей данного анализа выбросы не

будут удаляться, поскольку они могут считаться очень информативными, т.к. могут указывать на регионы, находящиеся в критическом состоянии и нуждающиеся в помощи.

Для выявлении зависимостей между индикаторами использовалась корреляция. Корреляция или корреляционная зависимость — статистическая взаимосвязь двух или более случайных величин при этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Иначе говоря она описывает взаимосвязь между изменениями двух переменных. Данный коэффициент принимать значения от -1 до 1.

Из-за различных показателей данных, варьирующихся в больших диапазонов, было использовано три метода масштабируемости: стандартизация, нормализация и стандартизация с анализом главных компонентов. Анализ главных компонентов (PCA) — это статистический метод, который используется для уменьшения размерности набора данных путем преобразования его в новое пространство переменных, называемых главными компонентами. Главные компоненты являются линейными комбинациями исходных переменных и ранжируются по убыванию вклада в объяснение вариации в данных. Этот метод позволяет определить индикаторы, представив их в виде новых переменных, которые обладают наибольшей изменчивостью.

В анализе главных компонент (PCA) стремятся сохранить основные различия между данными, используя пространство с более низкой размерностью. Это достигается путем создания новых компонент, которые являются линейными комбинациями исходных переменных и имеют меньшее количество компонент по сравнению с исходными переменными. Чтобы определить новую размерность используют схему осыпи.

Алгоритм K-средних – это простейший неконтролируемый алгоритм обучения, который решает проблему кластеризации. Этот алгоритм разбивает n наблюдений на k кластеров, где каждое наблюдение принадлежит кластеру с ближайшим средним, служащим прототипом кластера.

После нахождения наилучшего количества кластеров, при помощи метода локтя и метода силуэта, было выявлено три наиболее подходящих центроида для разбиения и дальнейшего анализа данных. Применение модели кластеризации в данном случае не выявило закономерностей, оно лишь под-

твердило общие знания. Кластеризацию можно рассматривать как этап предварительной обработки.

Дальнейшим шагом является построение линейной регрессии. Линейная регрессия — это математическая модель, описывающая связь нескольких переменных. Такая модели представляют собой статистическую процедуру, которая прогнозирует будущие события или показатели. Цель линейной регрессии заключается в создании математической модели, которая наилучшим образом описывает связь между переменными на основе имеющихся данных. Это позволяет предсказывать значения зависимой переменной на основе известных значений независимых переменных. Она основана на принципе минимизации суммы квадратов разностей между наблюдаемой зависимой переменной и предсказанными значениями, полученными с помощью линейного уравнения регрессии. Для наглядности и обоснованности наблюдений были построены остаточные графики и таблицы основных сведений линейной регрессии по имеющимся индикаторам и главному показателю "Индекс бедности".

Дальнейшим шагом является построение многомерной регрессии. Многомерная регрессия - это статистический метод анализа, который исследует связь между зависимой переменной и двумя или более независимыми переменными. В этом методе предполагается, что зависимая переменная зависит линейно от независимых переменных.

Один из основных подходов к многомерной регрессии - это метод наименьших квадратов (МНК). Он используется для оценки параметров регрессии, таких как коэффициенты наклона и свободный член, чтобы определить линейную зависимость между переменными. Цель МНК - минимизировать сумму квадратов остатков, то есть разницу между фактическими значениями зависимой переменной и значениями, предсказанными моделью.

Многомерная регрессия является мощным инструментом для анализа множества переменных и их влияния на зависимую переменную. В практической части используется многомерная регрессия с характеристиками исходного набора данных, без особенностей, а также с характеристиками с наибольшим значением R-квадрат. Результаты данного обучения представлены в таблице 1.

Таблица 1 – Результат многомерной регрессии с учетом различных параметров

Многомерная регрессия		**Оценка точности обучения**
Со всеми характеристиками исходного набора данных	R-квадрат	0.884
	скорректированный R-квадрат:	0.868
Без особенностей с мультиколлинеарностью и гетероскедастичностью		0.378
С характеристиками с наибольшим значением R-квадрат, найденным по линейной регрессии		0.829

При всех характеристиках исходного набора данных скорректированное значение R-квадрат составляет почти 90%, что считается хорошим результатом, но необходимо учитывать, что в него включены характеристики, которые показывают мультиколлиарность и гетероскедастичность. Это означает, что модель, возможно, не была хорошо подобрана. Это происходит потому, что каждый раз, когда добавляется признак в модель, R-квадрат увеличивается, даже если это происходит случайно. Следовательно, может показаться, что модель с большим количеством признаков подходит лучше просто потому, что в ней их больше, но это не обязательно означает, что это лучший выбор признаков.

После выделения зависимостей и разбиение субъектов Российской Федерации на кластеры, и анализа этих кластеров в них были выделены регионы в которых показатели отклоняются от запланированных. В таблице 2 представлены эти регионы.

Таблица 2 – Регионы, нуждающиеся в изменении показателей социально-экономического развития людей

	Регионы
Кластер 1	Республика Дагестан, Ростовская область, Чеченская республика, Ставропольский край
Кластер 2	Тюменская область, республика Саха (Якутия)

Реализация построенной модели выделила регионы, в которых уровень бедности принадлежит критической зоне, нуждается в корректировке и в корректировке факторов, влияющих на него. Для выработки более чётких рекомендаций необходимо провести дальнейший анализ набора данных, а именно добавить дополнительные признаки связанные с контекстом и ограничениями, ввести в модель системные проблемы такие как: коррупция, кризис политического/гражданского общества, природные катастрофы и другие риски.

ЗАКЛЮЧЕНИЕ

Экономическое неравенство продолжает оставаться одной из важнейших глобальных проблем. Бедность — это состояние, при котором человеку не хватает средств для удовлетворения своих основных потребностей, таких как пища, кров и одежда. И хотя она существует в каждой стране, в одних бедность более выражена, чем в других. В последние годы сначала пандемия, а затем и военный конфликт на Украине обострили проблему нехватки доходов и их неравенства в мировом масштабе. Многие страны включились в активную борьбу с бедностью, в том числе и Российская Федерация. Для этого была сформирована государственная программа, основные положения которой были озвучены в Едином плане по достижению национальных целей развития Российской Федерации на период до 2024 года и на плановый период до 2030 года от 1 октября 2021 года.

Целью дипломной работы является получение инструмента для анализа показателя бедности и его индикаторов для регионов Российской Федерации, выделения регионов с индикаторами, которые отличаются от значений в Едином плане.

Для достижения цели выбран набор данных, проведен необходимый статистический анализ, построена, обучена и реализована модель зависимости индекса бедности от других социально-экономических факторов. С помощью различных критериев проверены статистические характеристики, визуализированы результаты корреляционного и регрессионного анализов, сделан выбор индикаторов, сильно влияющих на индекс бедности. Это позволило выделить регионы, в которых индекс бедности и значения его индикаторов отличаются от значений Единого плана, что может выступать причиной корректировки региональных программ развития и выработать пути достижения цели сохранения населения, здоровья и благополучия людей.

Основные источники информации:

1. Garrett Grolemund, Hadley Wickham R for Data Science: Import, Tidy, Transform, Visualize, and Model Data . - 1st Edition изд. - Sebastopol: O'Reilly Media, 2017. - 518 с.
2. Международная организация по стандартизации // Википедия [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Международная_организация_по_стандартизации (дата обращения: 15.04.2023).
3. Международная электротехническая комиссия // Википедия [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Международная_электротехническая_комиссия (дата обращения: 15.04.2023).
4. Wes McKinney Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter . - 3rd Edition изд. - Sebastopol: O'Reilly Media, 2022. - 579 с.
5. Портал открытых данных Российской Федерации - Каталог государственных сайтов // Открытые данные России URL: https://gosbar.gosuslugi.ru/ru/sites/14/ (дата обращения: 10.05.2023).
6. Site Maintenance // Инфраструктура научно-исследовательских данных [Электронный ресурс] URL: https://data-in.ru/maintenance/ (дата обращения: 10.05.2023).
7. Федеральная служба государственной статистики // Федеральная служба государственной статистики (Росстат) [Электронный ресурс] URL: https://rosstat.gov.ru/ (дата обращения: 09.05.2023).
8. ЕМИСС // Единая межведомственная информационно-статистическая система [Электронный ресурс] URL: https://www.fedstat.ru/ (дата обращения: 10.05.2023).
9. Доходы населения. Опыт количественных измерений / А.Е. Суринов. - Москва : Финансы и статистика, 2000. - 428 с.
10. Большие данные // Википедия [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Большие_данные (дата обращения: 12.02.2023).
11. ГОСТ Р ИСО/МЭК "Информационные технологии. Большие данные. Обзор и словарь"от 01.11.2021 № 20546-2021 № 35.020. - Ст. 16
12. Минкомсвязь России готовится к регулированию больших данных // Адвокатская газета [Электронный ресурс] URL: https://www.advgazeta.ru/novosti/minkomsvyaz-rossii-gotovitsya-k-regulirovaniyu-bolshikh-dannykh/

(дата обращения: 03.03.2023).
13. Нейронная сеть // Википедия [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Нейронная_сеть (дата обращения: 03.03.2023).
14. НОУ ИНТУИТ | Лекция | Что такое Data Mining? // Национальный Открытый Институт ИНТУИТ [Электронный ресурс] URL: https://intuit.ru/studies/courses/6/6/lecture/158?page=2 (дата обращения: 05.03.2023).
15. Jeff Howe Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business . - 1st edition изд. - Strawberry Hills: Currency, 2008. - 322 с.
16. Обработка больших данных: основные методы // GeekBrains [Электронный ресурс] URL: https://gb.ru/blog/obrabotka-bolshikh-dannykh/ (дата обращения: 27.03.2023).
17. Многомерные статистические методы для экономистов и менеджеров : учеб. для студентов экон. спец. вузов / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. - Москва : Финансы и статистика, 1998. - 350 с.
18. Имитационное моделирование : учеб. пособие / М. С. Эльберг, Н. С. Цыганков. – Красноярск : Сиб. федер. ун-т, 2017. – 128 с.
19. Революция, которая изменит то, как мы живем, работаем и мыслим / В. Майер-Шенбергер, К. Кукьер. М.: Манн, Иванов и Фербер, 2014. 240 с.
20. Авдеева И.Л. Анализ зарубежного опыта использования глобальных технологий «BigData» // Интернетжурнал «НАУКОВЕДЕНИЕ» Том 8, №6 (2016) http://naukovedenie.ru/PDF/13EVN616.pdf (доступ свободный).
21. Указ Президента Российской Федерации "Единый план по достижению национальных целей развития Российской Федерации на период до 2024 года и на плановый период до 2030 года"от 21.07.2020 г. № 474 // Российская газета. - 2021 г.
22. Введение в машинное обучение с помощью Python : руководство для специалистов по работе с данными : Андреас Мюллер, Сара Гвидо ; [перевод с английского и редакция А. В. Груздева]. - Москвas : Диалектика, 2017. - 472
23. Корреляция // Википедия [Электронный ресурс] URL:Корреляция // Википедия URL: https://ru.wikipedia.org/wiki/Корреляция (дата обра-

щения: 11.03.2023).
24. I.T. Jolliffe Principal Component Analysis (Springer Series in Statistics). - 2nd Edition изд. - New York City: Springer, 2002. - 518 с.
25. Congming Shi, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, Jialei Liu A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm // EURASIP Journal on Wireless Communications and Networking. - 2021. - №31. - С. 1-16.
26. D'Agostino's K-squared test // Wikipedia [Электронный ресурс] URL: https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test (дата обращения: 23.04.2023).
27. Interpreting linear regression summary from statsmodels // ntegrab?? ifferenti??s [Электронный ресурс] URL: https://www.adrian.idv.hk/2021-07-16-statsmodels/ (дата обращения: 22.04.2023).
28. Linear regression diagnostics in Python | Jan Kirenz // Jan Kirenz URL: https://www.kirenz.com/post/2021-11-14-linear-regression-diagnostics-in-python/linear-regression-diagnostics-in-python/ (дата обращения: 24.04.2023).

29. Damodar N Gujarati, Dawn C. Porter Basic Econometrics. - 5th Edition изд. - New York City: McGraw-Hill Education, 2008. - 944 с.
30. Douglas C. Montgomery, Elizabeth A. Peck , G. Geoffrey Vining Introduction to Linear Regression Analysis. - 4th Edition изд. - Hoboken: Wiley-Interscience, 2006. - 640 с.