Оглавление

Введение

В настоящий момент на рынке происходит бум машинного обучения, тысячи компаний по всему миру видят в нём огромный потенциал и вкладывают деньги в данное направление. Его методы применяются повсеместно, начиная от распознования текста и диагностики заболеваний сердца по видеопотоку с веб-камеры, и заканчивая умными автомобилями, способными обходиться вез водителя. О востребованности специалистов и наработок в этой области говорит тот факт, что за исследование и анализ данных компании готовы платить огромные деньги [43].

Нарастающая потребность в исследовании процессов и показателей методами машинного обучения особенно заметна в финансовых организациях (кредитный скоринг, предсказание котировок акций) и государственных учреждениях с целью планирования деятельности и максимально эффективного распределения ресурсов. Время экспертных оценок осталось в прошлом. Конечно, так называемое «чутьё» всё ещё может принести какие-то дивиденды, но ни одна уважающая себя организация не станет целиком полагаться на иррациональную интуицию, когда в её распоряжении имеются инструменты для извлечения объективной информации из данных и построения на их основе выводов, точность которых поддаётся математической оценке.

В данной работе огромный потенциал методов машинного обучения используется для построения многомерных моделей таких социально-экономических показателей регионов России, как число территориальных организаций самоуправления на десять тысяч человек, общественное обсуждение законопроектов в сети Интернет и контроль за ходом государственных закупок по ряду институциональных, инфраструктурных и ресурсных факторов.

Объектом исследования являются органы местного самоуправления регионов Российской Федерации, осуществляющие контроль и планирование на основе социально-экономических показателей региона.

Предметом исследования являются методы оценки и прогнозирования социально-экономических индикаторов.

Целью данной работы является построение многомерных моделей классификации и регрессии ряда социально-экономических показателей регионов России.

Основные задачи:

1) Изучить теоретические аспекты исследуемых социально-экономических показателей;

2) Рассмотреть методы машинного обучения, потенциально применимые для построения многомерных моделей зависимостей;

3) Определить оптимальные методы машинного обучения для задачи обнаружения зависимостей в имеющихся данных;

4) Построить многомерные модели регрессии и классификации социально-экономических показателей на имеющихся наборах факторов;

5) Сделать заключение о возможности построения моделей зависимостей хорошего качества на ограниченных наборах наблюдений в данном конкретном случае и в общем, подвести итоги по полученным моделям.

Научная новизна работы заключается в использовании современных и эффективных инструментов при моделировании слабоизученных региональных экономических индикаторов, подробном и последовательном изложении выполняемых на каждом этапе исследования шагов, которое может быть использовано в качестве практического руководства при изучении других экономических вопросов.

В первой главе работы рассмотрены изучаемые социально-экономические показатели, представлены последние наработки по моделированию экономических величин.

Во второй главе представлены алгоритмы и методы машинного обучения, расположенные в естественном порядке нарастающей сложности лежащих в их основе математических выводов и заключений, от простейших линейных моделей до ARM регрессии, базирующейся на байесовском подходе к

определению весов коэффициентов. Тем не менее, метод главных компонент вынесен в конец лишь по причине того, что данный метод понижения размерности несколько выделяется на фоне методов математического моделирования.

В третьей главе на практике с использованием современных инструментов производится попытка построения моделей многомерной зависимости социально-экономических показателей от институциональных, инфраструктрных и ресурсных факторов, делается заключение по результатам этой работы.

В заключении подводятся итоги работы, оцениваются достигнутые результаты и их применимость на практике.

В работе используются наиболее широко распространённые в русскоязычной среде экономические и математические термины, большинство математических обозначений описано на первых страницах второй главы, до начала работы с ними.

1 Теоретический обзор исследуемых показателей

1.1 Территориальной общественное самоуправление

В условиях реформирования местного самоуправления одной из важнейших задач, стоящих перед органами местной власти в муниципальных образованиях, особенно на уровне крупных городов, является создание механизмов вовлечения жителей в решение местных вопросов. Из всего многообразия организационных форм участия населения в местном самоуправлении важная роль принадлежит территориальному общественному самоуправлению как социально-экономическому институту самоорганизации граждан по месту их жительства на части территории поселения для самостоятельного и под свою ответственность осуществления собственных инициатив по вопросам местного значения [19, с.1].

Федеральный закон «Об общих принципах организации местного самоуправления в Российской Федерации» дал четкое определение понятию «территориальное общественное самоуправление» и определил порядок правового регулирования создания и деятельности ТОС. Принятие закона стимулировало их образование. Они растут количественно и качественно. Появляются ассоциации и союзы, институты поддержки, как со стороны муниципальной власти, так и со стороны третьего сектора. Идет сложный процесс поисков взаимодействия между новыми организациями жителей и муниципалитетами. Этот многообразный опыт активно тиражируется, зачастую без критического осмысления и анализа.

В последние годы, по мнению большинства учёных и специалистов в области муниципального управления, территориальное общественное самоуправление становится наиболее массовой и эффективной формой непосредственного вовлечения граждан в процесс решения местных вопросов и рассматривается в качестве самостоятельного элемента системы организации местного самоуправления в муниципальных образованиях [20, с.1].

Опыт практической деятельности ТОС показывает, что данная форма самоорганизации жителей начинает оказывать всё большее влияние на

социальные процессы, происходящие внутри муниципальных территорий []. Особенно заметную роль стали играть органы ТОС в благоустройстве дворовых и уличных территорий, контроле за состоянием жилого фонда, организации досуга, защиты прав и интересов своих жителей [].

Заметный количественный рост органов ТОС в муниципальных образованиях и их практическое участие в управлении процессами жизнедеятельности людей вызывают потребность в получении объективной информации об их деятельности, формах и методах работы с населением. По мнению В. П. Максимова [19, с.2], круг всех заинтересованных управленческих структур и организаций можно разделить на следующие условные группы:

- органы государственного и муниципального управления, ассоциации муниципального сотрудничества, включая муниципальные мероприятия и учреждения;

- жители и актив ТОС, проживающие на территории, где функционируют созданные органы ТОС;

- органы ТОС в целях проведения сравнительного анализа результатов своей деятельности с другими организациями и органами территориального общественного самоуправления;

- бизнес-структуры, оказывающие спонсорскую помощь органам ТОС в реализации социально-экономических проектов;

- политические образования для взаимодействия с органами ТОС для объединения усилий по созданию гражданского общества.

В. П. Ляхов считает, что необходимость территориального общественного самоуправления в условиях модернизации России и реализации курса на демократизацию общественно-политических отношений не вызывает сомнений и обусловлена потребностью в укреплении реализации свобод граждан в процессе формирования самостоятельных инициатив в решении социально-экономических вопросов развития территории проживания граждан [17, с.1]. Поскольку Россия — страна регионов, анализировать работу ТОС следует начинать с регионального опыта.

Вместе с тем реальное развитие территориального общественного самоуправления в России поставило ряд проблем и идеологического и организационно-управленческого плана. То, как они будут решены, во многом определит перспективы развития ТОС, его место и роль в системе местного самоуправления.

Речь идет, прежде всего, о правовом поле ТОС. С одной стороны, еще не во всех муниципальных образованиях создана нормативно-правовая база. С другой стороны, нередки случаи наделения органов ТОС властными полномочиями, утверждения или назначения руководителей ТОС, и даже включение их в штат органов местного самоуправления. Такой подход присущ и некоторым «лидерам» ТОС, требующим от местных администраций зарплаты, удостоверения и обвиняющим муниципалитеты в нежелании поделиться властью. Во многих российских городах и поселениях сохранились с советских времен, и составляют основу самоуправления старые локальные организации, такие, как уличные комитеты, домкомы и т.д., деятельность которых исчерпывающим образом регламентирована нормативно-правовыми актами органов местного самоуправления. В ряде случаев неоправданно завышены нормы представительства на общих собраниях при выборах органов ТОС. Таким образом, выхолащивается сущность ТОС как самодеятельной организации, побуждающей и реализующей гражданские инициативы на локальной территории под свою ответственность.

Однако, по данным Э. С. Коложвари, ситуация меняется [15, с.2]. Он ссылается на призыв заместителя руководителя Администрации Президента к руководителям городов передавать часть функций органов местной власти социально ориентированным некоммерческим организациям. Однако, высказанная им идея об использовании ТОС для связывания деструктивной протестной деятельности [15, с.2] вызывает некоторое недоумение.

Интересно, что органы ТОС на своей территории могут играть разные роли или несколько ролей одновременно, как показано на рисунке 1.

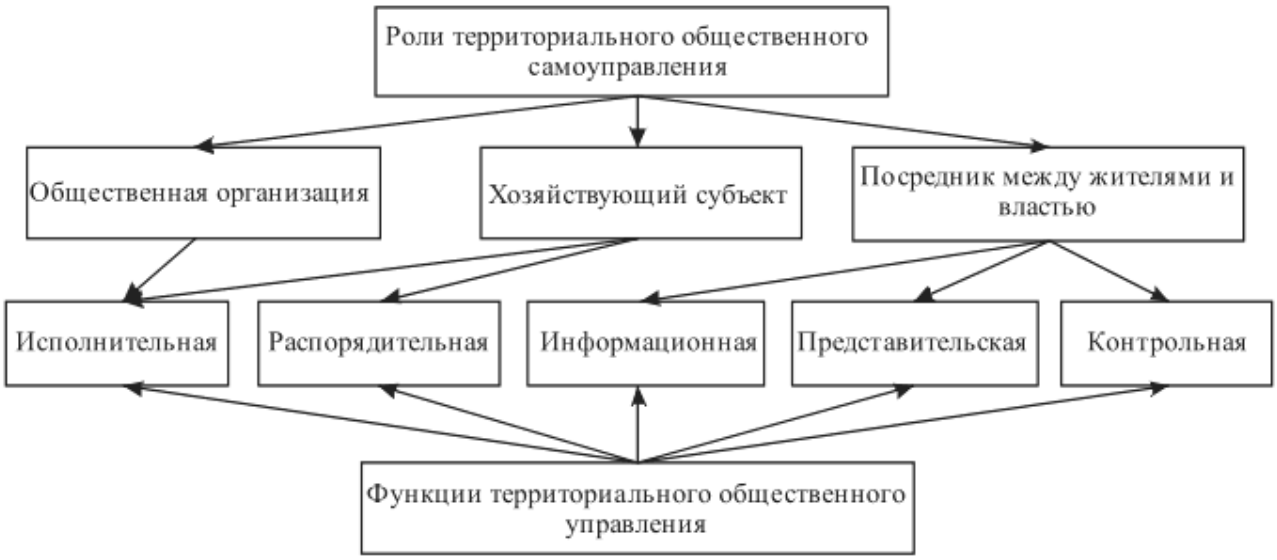

Рисунок 1 — Роли и функции территориального общественного самоуправления

Источник: по материалам [15]

По мнению В. П. Максимова, наиболее существенными признаками территориального общественного самоуправления, позволяющими выделить его доминирующее значение среди других форм вовлечения жителей в управленческие процессы в рамках муниципальных образований, являются [18, с. 2]:

- территориальный признак, обозначающий действия ТОС на определённой части территории муниципального образования;

- инициативный признак, характеризующий возникновение ТОС исключительно по желанию и инициативе самих жителей, проживающих на соответствующей территории;

- нормативно-правовой признак, устанавливающий правовой статус и место ТОС в системе местного самоуправления и являющийся субъектом муниципального права;

- функциональный признак, означающий реализацию функций, направленных на предоставление особого вида услуг по реализации общественных потребностей населения, защиты их прав и интересов.

По сведениям того же В. П. Максимова, доказательством важности и широкой распространённости ТОС, является тот факт, что уже по состоянию на 2003 год в городах Урало-Сибирской зоны органы ТОС охватывали от двадцати до девяноста двух процентов жителей территории, о чём свидетельствует рисунок 2. С этим трудно не согласиться.

Город	Количество комитетов (советов) ТОС	Численность жителей, тыс. чел.	Уровень охвата жилой территории города ТОС, %
Екатеринбург	23	1293,0	20,0
Ижевск	23	650,3	30,0
Новосибирск	80	1425,6	45,0
Омск	71	1133,9	72,0
Оренбург	48	516,6	50,0
Пермь	67	1000,1	58,6
Тюмень	28	500,2	42,0
Челябинск	119	1078,3	92,0

Рисунок 2 — количественный состав ТОС и их охват жилой территории в городах Урало-Сибирской зоны в 2003 году.

Источник: по материалам [18]

Всё вышеизложенное по ТОС свидетельствует о большой востребованности исследований такого социально-экономического показателя, как число организаций территориального общественного самоуправления в регионах страны.

1.2 Государственные закупки

В России, к настоящему времени, сформирована достаточно полная нормативно — законодательная база по организации процессов закупок продукции (услуг) для государственных нужд, включая конкурсные торги различных видов. Состав указанных нормативно-правовых актов включает большое количество (более 300) документов, как на федеральном уровне, так и на уровне отдельных субъектов Российской Федерации и органов местного самоуправления. В некоторых из них, по предварительным оценкам,

нормативная база в области регулирования государственных закупок является даже более проработанной, чем на федеральном уровне.

Законодательно-нормативная база достаточно подробно регулирует гражданско-правовые вопросы государственных закупок, взаимные права и обязанности сторон, состав и виды формируемых документов на различных этапах заключения контрактов и их исполнения.

Однако в федеральном законодательстве РФ отсутствует конкретное определение государственных и муниципальных заказов.

Существует несколько определений к понятию государственный заказ:

- государственный заказ — выдаваемый государственными органами и оплачиваемый из средств государственного бюджета заказ на изготовление продукции, выпуск товаров, проведение работ, в которых заинтересовано государство;

- государственный заказ — совокупность заключенных государственных контрактов на поставку товаров, производство работ, оказание услуг за счет средств государственного бюджета;

- государственный заказ — предложение, даваемое уполномоченной государственной организацией другой организации-поставщику о поставке товаров, работ, услуг для федеральных и региональных государственных нужд [12].

Под муниципальным заказом, в свою очередь, понимают заказ со стороны органов местного самоуправления и уполномоченных ими муниципальных учреждений на поставки товаров, выполнение работ и оказание услуг, связанных с решением вопросов местного значения и осуществлением отдельных государственных полномочий, переданных органам местного самоуправления федеральными законами и законами субъектов РФ.

В соответствии с Федеральным законом № 94-ФЗ:

- Под государственными нуждами понимаются обеспечиваемые за счет средств федерального бюджета или бюджетов субъектов Российской Федерации и внебюджетных источников финансирования потребности Российской

Федерации, государственных заказчиков в товарах, работах, услугах, необходимых для осуществления функций и полномочий Российской Федерации, государственных заказчиков либо потребности субъектов Российской Федерации, государственных заказчиков в товарах, работах, услугах, необходимых для осуществления функций и полномочий субъектов Российской Федерации, государственных заказчиков.

- Под муниципальными нуждами понимаются обеспечиваемые за счет средств местных бюджетов и внебюджетных источников финансирования потребности муниципальных образований, муниципальных заказчиков в товарах, работах, услугах, необходимых для решения вопросов местного малого бизнеса.

- Участниками размещения заказов являются лица, претендующие на заключение государственного или муниципального контракта. Участником размещения заказа может быть любое юридическое лицо независимо от организационно-правовой формы, формы собственности, места нахождения и места происхождения капитала или любое физическое лицо, в том числе индивидуальный предприниматель.

- Под государственным или муниципальным контрактом понимается договор, заключенный заказчиком от имени Российской Федерации, субъекта Российской Федерации или муниципального образования в целях обеспечения государственных или муниципальных нужд [5, с. 3].

Государственные и муниципальные закупки представляют собой крупный сегмент бюджетных расходов. В последние пару лет (по состоянию на 2012 год) ориентировочный объём госзакупок составил тринадцать триллионов рублей [27, с.1]. Но на пути постоянного развития необходимо не забывать опыт предыдущих лет. Можно выделить несколько этапов развития системы госзакупок в Российской Федерации [27, с.2]:

1) Период с 1992 по 1997 год. В этот период был принят ряд документов, призванный стать правовой базой регулирования госзакупок: 826-УП,, соответствующее постановление Правительства РФ о его реализации и 52-ФЗ.

Тем не менее, на этом этапе проведение торгов при размещении государственного заказа не являлось обязательным, что соответствующим образом использовалось экономическими агентами и требовало совершенствования правовой базы;

2) Период с 1997 по 2006 год. В данный промежуток времени были приняты правовые документы, которые должны были стать основой формирования современной системы регулирования государственныхзакупок: 305-УП и 97-ФЗ;

3) Период с 2006 по 2014 год. Начало очередного этапа становления отечественной системы государственных закупок, связанное с принятием 94-ФЗ. Он внёс множество изменений в процесс размещения государственного заказа.

4) Период с января 2014 года. В это время вступает в силу закон 44-ФЗ, появившийся по причине того, что комплекс проблем, накопившихся в российском госзаказе, уже невозможно было решить в рамках корректировки существовавших законов.

Несмотря на то, что закон о контрактной системе выступил в силу в 2014 году, говорить о том, что окончательный переход от старой системы госзаказа к новой свершился, было бы преждевременно. На данный момент ФЗ № 44 скорее выступает каркасом, который задает общие рамки функционирования системы закупок и нуждается в дальнейшей детализации, посредством нормативно-правовых актов.

Эксперты Высшей школы экономики (ВШЭ) полагают, что положительных результатов от внедрения контрактной системы стоит ждать не раньше 2017 года, когда будут сформированы основные институты новой системы государственных и муниципальных закупок. Потенциал экономии бюджетных средств от внедрения контрактной системы по их оценке может составить 700 млрд в год.

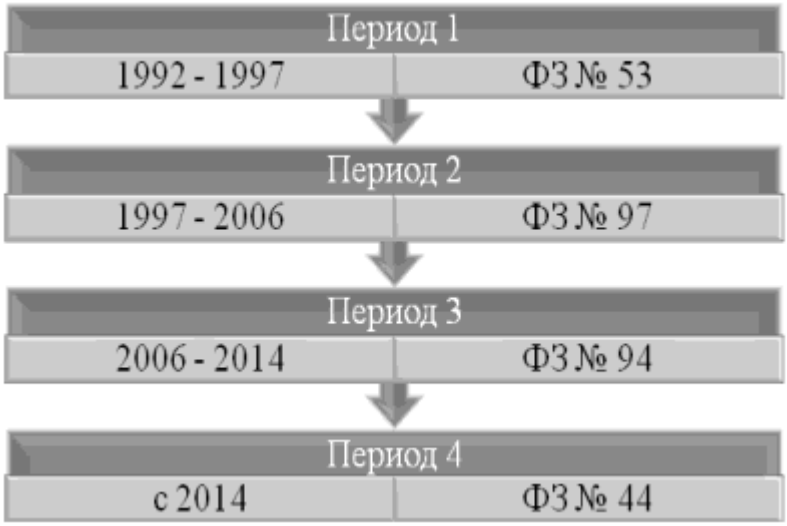

Рисунок 3 — Этапы госзакупок по Першину

Осуществление закупки обычно связано с выполнением определенной последовательности действий, которая в общем включает следующие укрупненные этапы:

- формирование заказа на закупку;

- проведение закупки;

- контроль за выполнением заключенных контрактов.

Формирование заказа на закупку обычно осуществляется в следующей последовательности:

1) формирование первоначального перечня закупаемых товаров, работ или услуг заказывающими подразделениями государственного или муниципального заказчика;

2) согласование проекта заказа с вышестоящими контролирующими и финансирующими органами государственного или муниципального заказчика, консолидация заказов на однотипную продукцию,

3) изменение номенклатуры или объемов закупаемой продукции;

4) утверждение заказа на поставку товаров, выполнение работ, оказание услуг.

После утверждения заказа на поставку продукции следует этап 2 (проведение закупки), который, как было показано выше, в основном осуществляется через проведение открытого конкурса.

Порядок проведения открытого конкурса обычно предусматривает следующие этапы:

1) проведение подготовительной работы перед проведением конкурса: уточнение и доработка заказа на поставку товаров (выполнение работ, оказание услуг), определение организатора конкурса;

2) разработка распорядительной и прочей документации, разработка конкурсной документации и извещения о проведении конкурса (2—4 недели);

3) публикация в печатных средствах массовой информации извещения о проведении открытого конкурса;

4) предоставление поставщикам по их запросам (в ряде случаев — после внесения соответствующей платы) конкурсной документации (обычно 1—2 недели после публикации извещения);

5) подготовка поставщиками своих предложений — конкурсных заявок (обычно 25—35 дней);

6) подача поставщиками и прием организатором конкурса конкурсных заявок (срок окончания приема заявок при проведении закупок для федеральных государственных нужд не может быть менее 45 дней со дня публикации извещения, в практике субъектов Российской Федерации обычно 30 дней);

7) публичная процедура вскрытия конкурсных заявок и оглашения их сути, по ходу процедуры ведется протокол;

8) оценка конкурсных заявок на предмет их соответствия требованиям конкурсной документации, соответствия поставщиков требованиям правоспособности и квалификации, соответствия технической стороны предложения требованиям заказчика, определения наиболее предпочтительной с коммерческой точки зрения конкурсной заявки (обычно 1—3 недели);

9) подведение итогов конкурса на заседании конкурсной комиссии, подписание протокола о результатах конкурса;

10) подготовка и подписание государственного (муниципального) контракта с выигравшим конкурс поставщиком (не позднее 20 дней со дня подведения итогов конкурса);

11) подготовка отчетности о результатах конкурса (1—2 недели).

Таким образом, в среднем процедура проведения открытого конкурса занимает 10—18 недель. Процедуры иных способов закупки (кроме двухэтапного конкурса) отличаются меньшими сроками, более простыми процедурами, менее сложными документами.

По результатам этапа закупки государственные или муниципальные заказчики имеют заключенный контракт, в ходе исполнения которого обеспечивается:

- контроль за своевременностью, комплектностью поставок (выполнения работ, оказания услуг);

- контроль качества поставляемых товаров (работ, услуг);

- контроль своевременности оплаты товаров (работ, услуг);

- рекламационная работа; подготовка отчетности.

Очевидно, что госзакупки крайней важны для экономики регионов и поэтому моделирование их зависимостей, к которому мы перейдём во в третьеё главе, крайне востребованно.

2 Алгоритмы и методы машинного обучения

2.1 Основные понятия и обозначения

Согласно российскому разделу Википедии, машинное обучение — это обширный подраздел искусственного интеллекта, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа, и извлекающая знания из данных [20]. Однако, данное определение практически не несёт смысловой нагрузки, сообщая читателю лишь то, что исследуемое понятие относится к математике.

Англоязычная версия ресурса [38] значительно более содержательна и предлагает два классических определения. Первое, предложенное Атуром Самуэлем в 1959 году, звучит следующим образом: машинное обучение — это область науки, занимающаяся исследованием способов обучения компьютерных систем без явного программирования результатов обучения [45]. Оно достаточно обще, но считается первым определением машинного обучения, первой попыткой выделить новое направление науки как самостоятельную и обособленную область исследований. Другое определение предложено Томом Митчеллом в 1997 году и носит более прикладное значение: говорят, что компьютерная программа обучается на опыте E по отношению к некоторому классу задач T и мере эффективности P, если её эффективность в задачах из T, измеренная с помощью P, увеличивается с накоплением опыта E [40]. Например, если взять классическую задачу А. Самуэля по игре в шашки, то за E принимают опыт игры большого количества партий в шашки, за T — задачу игры в шашки, а за P — вероятность победы программы в следующей партии. В данной работе мы в большей мере будем опираться на определение Т. Митчелла.

Все задачи машинного обучения могут быть отнесены к одному из двух обширных классов: обучению с учителем (также известном как обучение по прецедентам или обучение на размеченных данных) или обучению без учителя. Обучение по прецедентам имеет целью восстановление некоторой общей

зависимости по конечному числу известных примеров, представляющих собой пары объект-ответ, в то время как обучение без учителя предполагает поиск структуры в данных для выявления внутренних взаимосвязей, закономерностей и зависимостей, встречающихся в выборке, при условии, что дано только описание множества объектов. К обучению с учителем относятся задачи бинарной и многоклассовой классификации (распознавания образов), восстановления регрессии, ранжирования. Обучение с учителем преимущественно занимается проблемами кластеризации, визуализации и поиска аномалий. Как обучение с учителем, так и обучение без учителя широко используются в исследованиях экономических процессов и прикладных финансовых задачах. В рамках данной работы, в целях описания применённых далее методов при построении многомерных моделей социально-экономических показателей, будет рассмотрено преимущественно обучение с учителем.

Теперь, когда в общих чертах ясно, что представляет собой машинное обучение, перечислим основные понятия и обозначения, которые будут использоваться в работе и без которых введение новых определений не представляется возможным.

Объектом x из пространства объектов X называется элемент, для которого требуется сделать предсказание.

Ответом $y = y(x)$ из пространства допустимых ответов Y называется результат предсказания модели на объекте x .

Прецедентами (примерами) называются пары «объект-ответ» (x_I, y_i) .

Обучающей выборкой $X^l = (x_I, y_i)_{i=1}^l$ называется совокупность известных прецедентов, то есть множество объектов и соответствующих им ответов.

Тогда задачей обучения по прецедентам является восстановление зависимости целевой функции $y^* : X \to Y$ по обучающей выборке X^l путём построения решающей функции $a : X \to Y$, которая приближала бы $y^*(x)$ не только на объектах обучающей выборки, но и на всём множестве X . При этом

решающая функция a должна быть эффективно реализуема на компьютере, вследствие чего некоторые авторы называют её алгоритмом.

Признаковым описанием объекта называется совокупность его признаков (d -мерный вектор): $x=(x^1, x^2, ..., x^d)$, где x^i — признак, некоторая числовая характеристика, описывающая объект [37, с.. С формальной точки зрения признак — это отображение $f: X \to D$, где D_f — множество допустимых значений признака. Далее множество допустимых значений j -го признака будем обозначать D_j .

Совокупность признаковых описаний всех объектов выборки X^l , записанную в виде матрицы размера $l \times d$, называют матрицей «объекты-признаки»:

$$\hat{X} = \begin{pmatrix} x_{11} \dots x_{1d} \\ \dots \dots \dots \\ x_{l1} \dots x_{ld} \end{pmatrix} .$$

Вместе с ней обычно работают с вектором истинных ответов для всех объектов выборки $\hat{y}$:

$$\hat{y} = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix} .$$

В машинном обучении выделяют следующие виды признаков, каждый из которых необходимо обрабатывать и учитывать по-своему:

- бинарные признаки, $D_j = \{0, 1\}$;

- вещественные признаки, $D_j = \mathbb{R}$;

- категориальные признаки, D_j — неупорядоченное множество без отношения сравнения величины;

- порядковые признаки, D_j — упорядоченное множество;

- множественные признаки, D_j состоит из подмножеств некоторого упорядоченного множества.

При этом если все признаки признакового описания объектов имеют одинаковый тип, то исходные данные называются однородными, иначе — разнородными.

Функционалом ошибки $Q(a, X^l)$ называется некоторая характеристика качества работы алгоритма a на выборке X^l .

Тогда задача машинного обучения сводится к подбору такого алгоритма a из некоторого семейства A алгоритмов, для которого достигается минимум функционала ошибки.

Формализованное определение задачи обучения на размеченных данных звучит следующим образом. Дана обучающая выборка $X^l = (x_I, y_i)_{i=1}^l$, необходимо найти такой алгоритм $a \in A$, на котором будет достигаться минимум функционала ошибки:

$Q(a, X^l) \to \min\limits_{a \in A}$.

Конкретный тип задачи обучения с учителем зависит от множества возможных ответов Y :

- $Y = \{0,1\}$ соответствует задаче бинарной классификации;
- $Y = \{0,1,\dots,n\}$ соответствует задаче многоклассовой классификации;
- $Y = \mathbb{R}$, $Y = \mathbb{N}$ и любое другое множество допустимых ответов, содержащее бесконечное число элементов, соответствуют задаче восстановления регрессии.

2.2 Линейные модели

2.2.1 Модель линейной регрессии

Линейный алгоритм в случае регрессии выглядит следующим образом:

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x^j \quad ,$$

где w_0 — свободный коэффициент, x^j — признаки, а w_j — их веса.

Для представления формулы в боле компактном виде добавим $(d+1)$ -й признак, принимающий на каждом объекте значение 1. Тогда линейный алгоритм примет вид:

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = \langle w, x \rangle \quad ,$$

где $\langle w, x \rangle$ — скалярное произведение вектора весов на вектор признаков.

Мера ошибки должна быть выбрана таким образом, чтобы ее минимум достигался при правильном ответе. В качестве меры ошибки можно было бы взять модуль отклонения от прогноза $Q(a,y)=|a(x)-y|$, но функция модуля не является гладкой функцией, что вызывает сложности с оптимизацией градиентными методами. Поэтому мерой ошибки чаще всего выбирают квадрат отклонения:

$$(a(x)-y)^2 \quad .$$

При этом функционал ошибки, именуемый в данном случае среднеквадратичной ошибкой алгоритма, выглядит следующим образом:

$$Q(a,x)=\frac{1}{l}\sum_{i=1}^{l}(a(x_i)-y_i)^2 \quad .$$

Который для линейной модели принимает вид:

$$Q(w,x)=\frac{1}{l}\sum_{i=1}^{l}(\langle w,x_i\rangle-y_i)^2 \quad .$$

Таким образом, оценка качества линейной модели на обучающей выборке выглядит так:

$$Q(w,x)=\frac{1}{l}\sum_{i=1}^{l}(\langle w,x_i\rangle-y_i)^2\to\min_w \quad .$$

Среднеквадратичная ошибка может быть записана в матричном виде:

$$Q(w,\hat{X})=\frac{1}{l}\left\|\hat{X}w-\hat{y}\right\|^2\to\min_w \quad .$$

Можно найти аналитическое решение задачи минимизации:

$$w_*=(\hat{X}^T\hat{X})^{-1}\hat{X}^T\hat{y} \quad .$$

Однако, тогда придётся вычислять обратную матрицу, что требует порядка d^3 операций для d признаков, а в случае, когда матрица плохо обусловлена, численный метод нахождения обратной матрицы и вовсе не может быть применён. По этой причине обычно используют оптимизационный подход к решению.

Заметим, что среднеквадратическая ошибка является гладкой выпуклой функцией, что гарантирует существование лишь одного минимума и существование вектора градиента в каждой точке. Следовательно, возможно использования метода градиентного спуска. Для этого указывается некоторое

начальное приближение весов, после чего на каждой следующей итерации t из приближения w^{t-1} , полученного на предыдущей итерации, вычитается вектор градиента в соответствующей точке w^{t-1} , домноженный на коэффициент η_t , называемый шагом:

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, \hat{X}) \ .$$

Итерации останавливаются после достижения сходимости, то есть если разница двух последовательно полученных приближений меньше некоторого ε :

$$\left\| w^t - w^{t-1} \right\| < \varepsilon \ .$$

Пример работы метода градиентного спуска для случая парной регрессии можно увидеть на рисунке Ч. На случае парной регрессии, как на тривиальном, мы не будем останавливаться подробно. В случае же многомерной регрессии выражение для градиента в матричной форме имеет вид:

$$\nabla_w Q(w, \hat{X}) = \frac{2}{l} \hat{X}^T (\hat{X} w - \hat{y}) \ .$$

Тогда выражение для j -ой компоненты градиента будет содержать суммирование по всем компонентам обучающей выборки:

$$\frac{\partial Q}{\partial w_j} = \frac{2}{l} \sum_{i=1}^{l} x_i^j (\langle w, x_i \rangle - y_i) \ .$$

Очевидно, что такой подход неприемлем для экономических задач с большой обучающей выборкой, поскольку тогда даже одна итерация метода будет выполняться относительно продолжительное время. Поэтому на практике чаще всего используется метод стохастического градиентного спуска, основанный на том факте, что в формуле градиента в сумме каждое i -ое слагаемое для j -й компоненты показывает, как надо изменить вес w_j , чтобы улучшить качество для i -го объекта выборки. В стохастическом методе градиент функции качества вычисляется только на одном объекте обучающей выборки. Таким образом, после выбора начального приближения, веса на каждом следующем шаге вычисляются по формуле:

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, \{x_i\}) \ .$$

Заметим, что в стохастическом методе градиентного спуска ошибка хоть и уменьшается в среднем, но на каждом отдельном шаге может как увеличиваться, так и уменьшаться. Для сравнения на рисунке Ч представлены графики зависимости величины ошибки от номера итерации для градиентного спуска и стохастического градиентного спуска.

И всё же стохастический градиентный спуск обладает рядом существенных преимуществ:

- каждый шаг выполняется значительно быстрее шага обычного градиентного метода;

- не требуется постоянно хранить всю выборку в памяти машины;

- можно использовать для ситуаций, когда на каждом шаге известен только один элемент выборки, например, онлайн-обучения.

2.2.2 Линейный классификатор

Линейный классификатор похож на линейную регрессию, по сути он отличается лишь типом возвращаемого значения, для чего в формулу алгоритма добавляется операция взятия знака от выражения:

$$a(x)=sign\left(w_0+\sum_{j=1}^{d} w_j x^j\right) \quad .$$

Как и в случае линейной регрессии, добавлением ещё одного константного признака для каждого объекта можно привести формулу к более однородному виду:

$$a(x)=sign\left(\sum_{j=1}^{d+1} w_j x^j\right)=sign\langle w,x\rangle \quad .$$

Геометрический смысл линейного классификатора заключается в построении некоторой гиперплоскости в пространстве признаков таким образом, что объекты одного типа будут лежать по одну её сторону, а объекты другого типа — по другую. Но, согласно геометрическому смыслу скалярного произведения, расстояние от некоторого объекта x до гиперплоскости $\langle w,x\rangle=0$ равно: $\frac{|\langle w,x\rangle|}{\|w\|}$.

С этим связано понятие отступа. Отступ является величиной, определяющей корректность ответа. Если отступ M_i больше нуля, то классификатор даёт верный ответ для объекта x_i , в противном же случае — ошибается. Математически отступ записывается следующим образом:

$$M_i = y_i \langle w, x_i \rangle \ .$$

Естественным кажется определять качество алгоритма линейной классификации для обучающей выборки по доле неправильных ответов:

$$Q(a,x) = \frac{1}{l}\sum_{i=1}^{l} [a(x_i) \neq y_i] = \frac{1}{l}\sum_{i=1}^{l} [y_i \langle w, x_i \rangle < 0] = \frac{1}{l}\sum_{i=1}^{l} [M_i < 0] \ .$$

Выражение, стоящее под знаком суммы в последнем выражении, называется функцией потерь, в данной случае — пороговой. К сожалению, такая функция является разрывной в точке ноль, что не позволяет использовать метод градиентного спуска. Можно, конечно, использовать методы негладкой оптимизации, но они значительно сложнее в реализации. Поэтому на практике чаще всего используют гладкую оценку $[M_i < 0] \leq \widetilde{L}(M_i)$ пороговой функции и минимизируют не долю неправильных ответов, а некоторую функцию, являющуюся оценкой сверху:

$$Q(a,\hat{X}) \leq \widetilde{Q}(a,\hat{X}) = \frac{1}{l}\sum_{i=1}^{l} \widetilde{L}(M_i) \to \min_a \ .$$

Примерами таких функций потерь являются:

- логистическая функция потерь: $\widetilde{L}(M) = \log_2(\exp(-M))$;

- экспоненциальная функция потерь: $\widetilde{L}(M) = \exp(-M)$;

- кусочно-линейная функция потерь (используется в методе опорных векторов): $\widetilde{L}(M) = max(0, 1-M)$.

В случае логистической функции потерь функционал ошибки имеет вид:

$$\widetilde{Q}(a,\hat{X}) = \frac{1}{l}\sum_{i=1}^{l} \ln(\exp(-M)) = \frac{1}{l}\sum_{i=1}^{l} \ln(\exp(-y_i \langle w, x_i \rangle)) \ .$$

Данное выражение является гладким, то есть допустимо использование метода градиентного спуска.

Отметим также, что, даже если число ошибок равно нулю, в ходе обучения всё равно будут возрастать отступы — увеличиваться уверенность алгоритма в полученных результатах.

2.3 Проблема переобучения

Допустим, имеется некоторая экономическая величина и некоторый линейный классификатор, хорошо предсказывающий её значения на обучающей выборке. Пусть доля его ошибок на обучающей выборке составляет одну десятую от количества объектов. Это вполне приемлемая величина. Однако, вполне возможно, что для новой выборки доля ошибок в разы возрастёт. Такие ситуации, когда алгоритм вместо выявления закономерности самоподгоняется под обучающую выборку, называются переобучением.

К сожалению, выявить переобучение, имея лишь, обучающую выборку, невозможно, поскольку как хорошо обученный, так и переобученный алгоритмы покажут на ней хорошие результаты.

Существует несколько способов борьбы с переобучением:

- исключение части выборки из обучающей для последующей проверки на ней обученного алгоритма (отложенная выборка);

- кросс-валидация, несколько усложнённый вариант отложенной выборки;

- скользящий контроль по отдельным объектам (частный случай кросс-валидации с размером блоков разбиения в единицу);

- при оценке уровня обученности алгоритма использовать поправки на меру сложности его модели (регуляризация).

Одним из признаков переобученности модели являются большие веса при признаках. Они могут возникать как вследствие простого переусложнения модели, так и в результате наличия мультиколлинеарности, то есть линейной зависимости, между признаками. Допустим, действительно существуют такие коэффициенты $\alpha_1,\dots,\alpha_d$, что для любого x_i из выборки выполняется:

$$\alpha_! x_i^1+\dots+\alpha_d x_i^d=\langle \alpha, x_i\rangle=0 \quad .$$

Допустим, было найдено некоторое решение задачи оптимизации:

$$w_* = argmin_w \frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - y_i)^2 \quad .$$

Но тогда для вектор весов $w_1 = w_* + t\alpha$ получим другое решение задачи оптимизации: $\langle w_* + t\alpha, x\rangle = \langle w_*, x\rangle + t\langle \alpha, x\rangle = \langle w_*, x\rangle$. Оно будет также хорошо описывать данные, как и исходный алгоритм. Вообще говоря, в этом случае имеется бесконечное множество решений, при этом многие из них будут иметь большие веса перед признаками и далеко не все покажут хорошую обобщающую способность. Таким образом, мультиколлинеарность также способствует переобучению.

Для борьбы с большими весами используется регуляризация, то есть прибавление некоторого регуляризатора, домноженного на коэффициент регуляризации λ к значению функционала ошибки с последующей минимизацией не функционала ошибки, а полученной суммы. Например, для квадратичного регуляризатора, называемого также L_2 -регуляризатором, задача оптимизации примет вид:

$$Q(w, \hat{X}) + \lambda \|w\|^2 \to \min_w \quad ,$$

где $\|w\|^2 = \sum_{j=1}^{d} w_j^2$.

Как видно из этого уравнения, чем выше коэффициент регуляризации, тем ниже веса при признаках, а с ними и сложность модели. Однако, теперь исследователь встаёт перед дилеммой выбора оптимального значения λ , достаточно большого, чтобы не допустить переобучения, но при этом не зануляющего значимых коэффициентов и позволяющего обнаружить закономерности в данных.

Также широко применяется L_1 -регуляризатор, представляющий собой норму вектора весов:

$$\|w\|_1 = \sum_{j=1}^{d} |w_j| \quad .$$

Заметим, что, в отличии от L_2 -регуляризатора, L_1 -регуляризатор не является гладким. Кроме того, он обладает интересной способностью к

обнулению части весовых коэффициентов и может использоваться для отбора признаков.

Рассмотрим теперь немного подробнее процесс кросс-валидации. При использовании кросс-валидации выборка делится на k частей одинакового или примерно одинакового размера. После этого каждая из этих частей по очереди используется в качестве тестового набора, а все остальные — в качестве обучающей выборки. После того, как каждый блок побывает в качестве тестового, будет получено k показателей качества, а итоговая оценка получается путём их усреднения. При этом число блоков выбирается достаточно произвольно, стоит только учитывать, что при большом количестве блоков получаются ненадёжные, но несмещённые оценки, тогда как при малом количестве блоков оценки будут надёжными и смещёнными.

При разбиении данных на блоки для кросс-валидации не следует забывать про перемешивание объектов, поскольку нередко их порядок непроизволен и подчиняется, например, какому-то правилу хранения данных, а это может оказать существенное влияние на результаты моделирования. Единственное исключение — задачи предсказания, когда данные упорядочены по времени и их перемешивание недопустимо.

При обучении часть параметров алгоритма не может быть получена из обучающей выборки. Такие параметры называются гиперпараметрами. К ним относится параметр регуляризации λ в случае использования регуляризации и степень полиному в задаче регрессии с семейством алгоритмов, заданных множеством полиномов различных степеней.

2.4 Метрики качества в машинном обучении

Метрики качества применяются для:

- задания функционала ошибки на этапе обучения;
- подбора гиперпараметров при измерении качества на кросс-валидации;
- оценки итоговой модели.

Первая метрика уже упоминалась в предыдущих разделах. Это среднеквадратичная ошибка (MSE, mean square error):

$$MSE(a, \hat{X}) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \quad .$$

К ней применим метод градиентного спуска, но, к сожалению, из-за второй степени этот функционал сильно штрафует за большие ошибки, в результате чего штраф на выбросах будет очень значительным и сильно повлияет на результат. Таким образом, есть риск, что алгоритм будет настраиваться на аномальные объекты и построенная модель будет не слишком удачной.

Следующая метрика похожа на предыдущую. Это средняя абсолютная ошибка (MAE, mean absolute error):

$$MAE(a, \hat{X}) = \frac{1}{l}\sum_{i=1}^{l}|a(x_i) - y_i| \quad .$$

Данный функционал не так прост в обращении из-за наличия разрыва в нуле, но зато значительно более устойчив к выбросам вследствие своей линейности.

Ещё одной метрикой, связанной со среднеквадратичной ошибкой, является коэффициент детерминации. Этот коэффициент показывает, какую долю дисперсии в целевом векторе y модель сумела объяснить:

$$R^2(a, \hat{X}) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \bar{y})} \quad ,$$

где $\bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$.

Для разумных моделей коэффициент детерминации лежит в пределах от нуля до единицы, причём $R^2 = 1$ соответствует идеальной модели, $R^2 = 0$ говорит о качестве модели на уровне наилучшей константной. Он может принимать и отрицательные значения, но в таких случаях построенная модель крайне плоха.

До этого момента рассматривались лишь симметричные модели, но, вообще говоря, на практике во многих моделях величина штрафа за недопрогноз и перепрогноз не одинакова. Например, магазин, продающий оргтехнику, при заниженном прогнозе спроса потеряет прибыль и лояльность покупателей, а завышенный прогноз выльется лишь в относительно незначительное повышение затрат на хранение продукции на складе.

В таких случаях часто используется квантильная функция потерь (квантильная ошибка):

$$\rho_\tau(a,\hat{X})=\frac{1}{l}\sum_{i=1}^{l}((\tau-1)[y_i<a(x_i)]+\tau[y_i\geq a(x_i)])(y_i-a(x_i)) \quad .$$

Параметр $\tau\in[0,1]$ определяет, за что штрафовать сильнее — за недопрогноз или за перепрогноз. Если τ ближе к единице, то больший штраф будет назначаться за недопрогноз, а при близости к нулю — за перепрогноз.

Для задач классификации естественно выбрать в качестве метрики число неправильных ответов. Однако, в отличие от задач регрессии, где метрики, как правило, минимизируются, в задачах классификации принято выбирать метрики так, чтобы речь шла об их максимизации. Поэтому на практике используется количество правильных ответов (accuracy):

$$accuracy(a,\hat{X})=\frac{1}{l}\sum_{i=1}^{l}[a(x_i)=y_i] \quad .$$

Данная метрика относительно проста и потому широко используется, однако имеет ряд существенных недостатков.

Во-первых, она плохо ведёт себя на несбалансированных выборках. Рассмотрим выборку из 1000 объектов, из которых 950 принадлежат к классу +1 и 50 к классу -1. Тогда константный классификатор, относящий любой объект к +1, покажет долю правильных ответов в 0.95, что может создать ошибочное впечатление хорошей предсказательной силы алгоритма. Для того, чтобы не впасть в подобное заблуждение, необходимо учитывать соотношение классов в выборке и считать, что разумный алгоритм должен иметь $accuracy\in[q_0,1]$, где q_0 — число объектов в самом крупном классе.

Во-вторых, доля верных ответов не учитывает потенциально разные цены разных типов ошибок. Вследствие этого её применение в подобных задачах невозможно.

При работе с задачами классификациями удобно использовать так называемую матрицу ошибок — таблицу, в которой указано количество верных срабатываний (true positive), ложных срабатываний (false positive), ложных пропусков (false negative) и истинных пропусков (true negative) предсказаний алгоритма относительно истинных значений ответов на объектах выборки. Схема матрицы ошибок представлена в таблице 1.

Таблица 1 — схема матрицы ошибок

	$y=+1$	$y=-1$
$a(x)=+1$	True Positive (TP)	False Positive (FP)
$a(x)=-1$	False Negative (FN)	True Negative (TN)

На основе этой таблицы вводится ещё две метрики. Первая, точность (precision), показывает, насколько можно доверять классификатору в случае срабатывания:

$$precision(a,\hat{X})=\frac{TP}{(TP+FP)} \quad .$$

Другая метрика, называемая полнотой (recall), показывает, на какой доле объектов первого класса срабатывает алгоритм:

$$recall(a,\hat{X})=\frac{TP}{(TP+FN)} \quad .$$

Эти две метрики используются в, например, задаче кредитного скоринга. Допустим, банк хочет сохранять количество невозвратных кредитов на уровне пяти процентов. В таком случае задача превращается в задачу максимизации полноты при условии сохранения точности на уровне большем или равном девяноста пяти процентам. Другим примером может быть задача медицинской диагностики, то есть когда когда необходимо построить модель, определяющую, есть ли заболевание у пациента, при условии, что алгоритм

верно определит наличие болезни не менее чем у восьмидесяти процентов из действительно больных пациентов. Тогда говорят, что это задача максимизации точности при условии удержания уровня полноты большим или равным восьмидесяти процентам.

Существуют также задачи, в которых необходимо максимизировать как точность, так и полноту. Естественно, первым приходящим в голову способом это сделать является использование в качестве метрики среднего арифметического точности и полноты:

$$A = \frac{1}{2}(precision + recall) \ .$$

Однако, в случае среднего арифметического для константной модели будут получаться неоправданно высокие оценки качества алгоритма, что приведёт к тому, что константный и разумный алгоритмы будут лежать на одной линии уровня, а значит, мы не сможем их различить по значению метрики.

Для исправления данного недостатка среднего арифметического можно использовать в качестве метрики минимум от значений точности и полноты:

$$M = min(precision, recall) \ .$$

Но тогда придётся иметь дело с другой проблемой — при одинаковом значении минимального показателя, второй показатель не влияет на величину метрики, в результате чего существенно различающиеся по качеству предсказания алгоритмы будут лежать на одной линии уровня, а значит, неразличимы.

Из-за перечисленных недостатков среднего арифметического и минимума, в качестве метрики в задачах классификации чаще всего используют сглаженный минимум, называемый F -мерой:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \ .$$

Помимо отсутствия недостатков, характерных для двух перечисленных перед этим метрик, формула F -меры легко модифицируется для случая, когда один из параметров важнее другого:

$$F = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision + recall} \ .$$

Тогда при значениях β , больших единицы, предпочтение оказывается точности, а при значениях параметра, меньших одного — полноте.

Большое количество алгоритмов бинарной классификации сперва вычисляют вещественное число $b(x)$, являющееся оценкой вероятности принадлежности классу, а затем сравнивают его с некоторым порогом t :

$$a(x)=[b(x)>t] \quad .$$

Более того, обычно мы заинтересованы именно в нахождении значения оценки принадлежности, а порог выбирается позже в зависимости от того, что на важнее, точность или полнота.

Рассмотрим несколько способов получения оценки принадлежности классу.

Первый из них основан на использовании кривой точности-полноты или PR -кривой. В данном методе по оси абсцисс откладывается полнота, а по оси ординат — точность, при этом каждой точке кривой будет соответствовать классификатор с некоторым значением порога. Примеры PR -кривых для задач с небольшим и значительным количеством объектов представлены на рисунке Ч.

Заметим, что PR -кривая выходит всегда из начала координат, а заканчивается точкой $(1,r)$, где r — доля объектов положительного класса. Она строится в осях точности и полноты, следовательно, изменяется при изменении баланса классов. На практике обычно в качестве метрики используют площадь под этой кривой. Такая метрика носит название $AUC-PRC$ (area under curve — precision-recall curve).

Другим способом оценки принадлежности к классу $+1$ является ROC -кривая, на оси абсцисс которой откладывается FPR (false positive rate), а на оси ординат — TPR (true positive rate):

$$FPR=\frac{FP}{FP+TN} \quad , \qquad TPR=\frac{TP}{TP+FN} \quad .$$

Примеры ROC -кривых представлены на рисунке Ч. Заметим, что любая ROC -кривая исходит из точки $(0,0)$ и приходит в точку $(1,1)$, а идеальный классификатор пройдёт через точку $(0,1)$. Площадь под этой

кривой характеризует оценку вероятности того, что при случайном выборе объектов двух классов, положительный объект получит большую оценку принадлежности к положительному классу, чем отрицательный. Заметим также, что при изменении баланса классов и неизменных свойствах объектов величина данной метрики не изменяется.

2.5 Метод максимизации правдоподобия

Пусть случайная величина x имеет распределение $F(x,\theta)$, X^n — выборка размера n :

$$X \sim F(x,\theta) \quad , \quad X^n = (X_1, \dots, X_n) \quad .$$

Тогда функция правдоподобия имеет вид:

$$L(X^n,\lambda) = \prod_{i=1}^{n} P(X = X_i,\theta) \quad .$$

Поскольку при логарифмировании положение макисмумов функции не изменяется, удобнее работать с логарифмом правдоподобия, поскольку тогда мы перейдём от произведения к суммированию в формуле:

$$\ln L(X^n,\lambda) = \sum_{i=1}^{n} \ln P(X = X_i,\theta) \quad .$$

Тогда оценкой максимального правдоподобия называется величина:

$$\hat{\lambda}_{ОМП} = argmax_{\lambda} \ln L(X^n,\lambda) \quad .$$

Для непрерывной случайной величины метод максимального правдоподобия записывается аналогично:

$$X \sim F(x,\theta) \quad , \quad L(X^n,\lambda) = \prod_{i=1}^{n} f(X = X_i,\theta) \quad , \quad \hat{\lambda}_{ОМП} = argmax_{\lambda} L(X^n,\lambda) \quad .$$

Метод максимального правдоподобия обладает следующими свойствами:

- состоятельность, то есть получаемые оценки при увеличении объёма выборки начинают стремиться к истинным значениям:

$$\hat{\lambda}_{ОМП} \to 0 \quad \text{при} \quad n \to \infty \quad .$$

- асимптотическая нормальность, то есть с ростом объёма выборки оценки максимального правдоподобия всё лучше описываются нормальным

распределением со средним, равным истинному значению θ , и дисперсией, равной величине, обратной к информации Фишера:

$$\hat{\lambda_{ОМП}} \sim N(\theta, I^{-1}(\theta)) \text{ при } n \to \infty \ .$$

При решении задачи регрессии значение ответа можно описать в качестве суммы регрессионной функции и случайного шума:

$$y = a(x) + \varepsilon \ .$$

Если этот случайный шум имеет нормальное распределение с нулевым средним и дисперсией σ^2 , то задача минимизации среднеквадратичной ошибки даёт оценку максимального правдоподобия для регрессионной функции $a(x)$:

$$a_* = argmin_a \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2 \ .$$

Данный факт позволяет применять в задаче регрессии свойства метода максимального правдоподобия, например, определять значимость признаков и осуществлять их отбор или строить доверительные интервалы для значения отклика на новых объектах.

Допустим теперь, что a — константа, а y представляет собой случайную функцию с плотностью распределения $f(t)$. Тогда среднеквадратичная ошибка примет вид:

$$Q(a) = \int_t (a - t)^2 f(t)\, dt \ .$$

Но тогда можно показать, что наилучшая константа, аппроксимирующая значение отклика в смысле среднеквадратичной ошибки — это математическое ожидание:

$$a_* = argmin_a Q(a) = mean(y) \ .$$

Если же $a(x)$ — произвольная функция от признаков, то то функционал среднеквадратичной ошибки примет вид:

$$Q(a, \hat{X}) = \int_t (a(x) - t)^2 f(t)\, dt \ ,$$

а его минимум будет представлять собой условное математическое ожидание:

$$a_*(x) = argmin_a Q(a(x)) = mean(y|x) \ .$$

Таким образом, в случае конечной выборки:

$$Q(a(x),\hat{X})=\frac{1}{l}\sum_{i=1}^{l}(a(x_i)-y_i)^2 \ ,$$

а оценка, полученная при минимизации среднеквадратичной ошибки, будет лучшей аппроксимацией условного математического ожидания:

$$a_*(x)=argmin_a Q(a,\hat{X}) \ .$$

В случае линейной регрессии, когда отклик моделируется линейной комбинацией, наилучшей линейной аппроксимацией условного математического ожидания $mean(y|x)$ является выражение $\langle w_*, x_i\rangle$:

$$Q(w,\hat{X})=\frac{1}{l}\sum_{i=1}^{l}(\langle w,x_i\rangle-y_i)^2 \ , \qquad w_*=argmin_w Q(w,\hat{X}) \ .$$

Но оказывается, что условное математическое ожидание доставляет минимум не только среднеквадратичной ошибки, а более широкого круга функций потерь, называемых дивергенциями Бергмана и порождаемых любой непрерывной дифференцируемой выпуклой функцией φ:

$$Q(a,\hat{X})=\varphi(y)-\varphi(a(\hat{X}))-\varphi'(a(\hat{X}))(y-a(\hat{X})) \ .$$

Рассмотренная нами ранее средняя абсолютная ошибка не входит в семейство дивергенций Бергмана. При её минимизации получается оценка не условного математического ожидания, а оценка условной медианы:

$a_*=argmin_a Q(a,\hat{X})$ — лучшая аппроксимация $med(y|x)$.

Несимметричная абсолютная функция ошибки имеет как бы наклонённый график по сравнению с графиком симметричной абсолютной ошибки. При минимизации такого функционала получается лучшая оценка для соответствующего условного квантиля:

$a_*=argmin_a Q(a,\hat{X})$ — лучшая аппроксимация $y|x$ порядка τ .

Можно сказать, что математическое ожидание квадрата ошибки регрессии представляет собой сумму трёх компонент:

$$mean(a_*(x)-y)^2=(mean(a_*(x))-a(x))^2+dispersion(a_*(x))+\sigma^2 \ ,$$

где первая компонента — квадрат смещения, вторая — дисперсия оценки, а третья — шум.

Первые две компоненты зависят от выбора модели, третья же характеризует данные и потому от модели не зависит.

Метод наименьших квадратов даёт несмещённые оценки. Регуляризация же позволяет получать смещённые оценки с меньшим квадратом смещения за счёт уменьшения дисперсии.

В байесовой статистике гребневая регрессия (ridge, регрессия с L_2 -регуляризатором) соответствует заданию нормального априорного распределения на коэффициенты линейной модели, а метода лассо (lasso, регрессия с L_1 -регуляризатором) — заданию лапласовского априорного распределения.

2.6 Метод автоматического отбора значимых признаков

В основе метода автоматического отбора значимых признаков лежит принцип, являющийся аналогом бритвы Оккама для машинного обучения: при выборе из двух моделей, одинаково хорошо объясняющих данные, всегда стоит выбирать более простую [51].

ARD регрессия похожа на байесовскую версию гребневой регрессии. Она основывается на предположении, что каждый вес имеет собственную дисперсию[39]. Их ищут по максимуму апостериорной информации:

$$w_{MP} = argmax\, p(w|t, \hat{X}, A) = argmax\, p(t|\hat{X}, w)\, p(w|A) \ ,$$

где $p(w|A) = \frac{\sqrt{det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} w^T A w\right)$ — априорное распределение, играющее роль регулятора весов.

Действительно, при использовании ARD семейство матриц расширяется:

$$A = \{diag(\alpha_1, \ldots, \alpha_m) | \alpha_j \geq 0\} \ .$$

Таким образом, каждый вес получает индивидуальный коэффициент регуляризации, который определяет степень адаптируемости конкретного веса под исходные данные.

Метод автоматического отбора значимых признаков хорошо работает для небольших наборов данных, но достаточно сложен в вычислительном плане, и потому редко используется на выборках в десятки тысяч объектов.

2.7 Метод главных компонент

Метод главных компонент применяется к данным, записанным в виде матрицы X — прямоугольной таблицы чисел размерностью I строк и J столбцов. Обычно строки этой матрицы называются образцами, а столбцы — переменными.

Как правило, перед применением PCA данные автошкалируются, то есть к ним применяются операции центрирования и нормирования:

$$\tilde{x_{i,j}} = \frac{\left(x_{i,j} - \frac{\left(x_{1j} + \ldots + x_{Ij}\right)}{I}\right)}{\sqrt{\frac{\sum_{i=1}^{I} \left(x_{ij} - m_j\right)^2}{I}}} \quad .$$

Центрирование используется, поскольку оригинальная модель метода главных компонент не содержит свободного члена, а нормирование для выравнивания вклада разных переменных в модель. Однако в задачах, где структура исходных данных предполагает однородность и гомоскедастичность, подготовка данных не только не нужна, но и вредна [1, с. 217].

Итак, пусть имеется матрица переменных X размерностью $I \times J$, где I — число образцов, а J — количество независимых переменных (столбцов), которых, как правило, много ($I \gg 1$). В метода главных компонент используются новые формальные переменные t_a ($a = 1, \ldots, A$), являющиеся линейной комбинацией исходных переменных x_j ($j = 1, \ldots, J$):

$$t_a = p_{a1} x_1 + p_{aJ} x_J \quad .$$

С помощью этих новых переменных матрица X разлагается в произведение двух матриц T и P : $X = TP^T + E = \sum_{a=1}^{A} t_a p_a^T + E$.

Матрица T называется матрицей счетов (scores), матрица P называется матрицей нагрузок, а матрица E — матрицей остатков.

Новые переменные нзываются главными компонентами (principal components). Они ортогональны между собой и их число заведомо меньше числа исходных переменных, поэтому PCA в основном применяется для сжатия данных, а точнее — для смены системы координат на более подходящую для описания данных в пространстве меньшей размерности.

Главные гипотезы метода главных компонент:

- линейность (позволяет упростить переход к новому базису [46, с. 6]) ;

- большая вариация несёт интересующие нас сведения;

- главные компоненты ортогональны (позволяет реализовать метод главных компонент с помощью инструментария линейной алгебры [46, с. 8]).

Также следует помнить о том, что PCA непараметричен, а значит не налагает условий на способ измерения данных и даёт независимый от пользователя определённый результат для каждого их набора. С одной стороны, это, безусловно, плюс, так как позволяет использовать этот метод без особой подготовки. С другой стороны, это и слабость метода. К примеру, если у нас есть набор точек, характеризующих положение некоторого фрагмента колеса в пространстве, записанный в декартовых координатах, то PCA окажется бесполезен.

Теперь, описав все методы, которые мы будем применять для исследования социально-экономических показателей регионов России, мы можем перейти к практической части — построению моделей зависимостей.

3 Разработка моделей многомерной классификации и регрессии социально-экономических показателей России

3.1 Описание исходных социально-экономических данных

Современные методы машинного обучения позволяют строить модели зависимостей для точного предсказания целевых переменных по набору факторов. Это используется в совершенно разных областях науки и в огромном количестве практических задач. В экономике их применяют преимущественно для оценки значения показателей, которые обычно определяются методом экспертных оценок, или для выделения факторов, в наибольшей степени влияющих на индикатор [22, с. 215]. В данной работе предпринимается попытка построения моделей, позволяющих предсказывать (метка — описание):

- cancelled — контроль за ходом госзакупок (доля отменённых конкурсов в общем количестве);

- discussion — общественное обсуждение законопроектов в сети Интернет (бинарный признак);

- tsg — количество зарегистрированных организаций территориального общественного самоуправления на десять тысяч человек, штук.

Значения предсказываемых показателей и их базовые статистические характеристики приведены на рисунке 4.

	cancelled	public_discussion	tsg
index			
vo	2.564	0	7.747
ra	1.889	0	2.227
ao	3.579	1	0.735
rk	4.098	1	0.000
kk	27.273	1	12.835
ro	3.176	0	0.863

	cancelled	public_discussion	tsg
count	6.000000	6.000000	6.000000
mean	7.096500	0.500000	4.067833
std	9.914503	0.547723	5.132527
min	1.889000	0.000000	0.000000
25%	2.717000	0.000000	0.767000
50%	3.377500	0.500000	1.545000
75%	3.968250	1.000000	6.367000
max	27.273000	1.000000	12.835000

Рисунок 4 — целевые показатели и их статистические характеристики

В качестве факторов будут использоваться следующие индикаторы (представлены в формате метка — описание):

- Институционные факторы:

а) elections — конкурентность выборов (бинарный признак);

б) parliament — уровень парламентской конкуренции в процентах;

в) executive — уровень конкуренции при формировании исполнительной власти в процентах;

г) citizens_election — включенность граждан в избирательный процесс в процентах;

д) citizens_org — участие граждан в деятельности общественных организаций в процентах;

- Инфраструктурные факторы :

а) internet_all — количество абонентов в сети Интернет в тысячах;

б) internet_mobile — количество абонентов мобильного интернета на сто человек в единицах;

в) internet_pc — число персональных компьютеров на сто человек в единицах;

г) org_access — организации, связанные с бизнесом и использующие интернет, от общего числа обследованных организаций в процентах;

д) org_site — организации, связанные с бизнесом и имеющие веб-сайт, от общего числа обследованных организаций в процентах;

е) org_pc — число персональных компьютеров на сто работников в бизнесе в единицах;

ж) edm — наличие систем электронного документооборота от общего числа организаций в процентах;

з) edm_external — автоматический обмен данными между своими и внешними информационными системами от общего числа обследованных организаций в процентах;

и) authority_access — организации, имеющие отношение к органам власти и использующие интернет, от общего числа обследованных организаций в процентах;

к) public_services — доступность государственных услуг, в том числе за счёт сокращения сроков предоставления (бинарный признак);

л) open_data — наличие инфраструктуры открытых данных, в том числе государственных (бинарный признак);

м) open_election — открытость процесса выборов (бинарный признак);

- Ресурсные факторы:

а) average_edu — доля населения со средним образованием в процентах;

б) high_edu — доля населения с высшим образованием в процентах;

в) degree — доля населения с учёными степенями в процентах;

г) ict — доля специалистов в ИКТ в процентах;

д) grp — ВРП на душу населения в рублях;

е) income — среднедушевые доходы населения в рублях;

ж) po — объём использованного программного обеспечения в процентах;

з) invest — удельный вес инвестиций в основной капитал в ВВП в процентах;

и) venture — доступность венчурного капитала в единицах;

к) pc — количество персональных компьютеров на 100 человек в единицах;

л) nt — затраты организаций на сетевые технологии в миллионах рублей;

м) ict_grp — удельный вес затрат на ИКТ в ВРП в процентах.

Для регионов Российской Федерации в данных используются следующие обозначения: vo — Волгоградская область, ra — Республика Адыгея, ao —

Астраханская область, rk — Республика Калмыкия, kk — Краснодарский край, ro — Ростовская область.

Необходимо отметить, что в работе использовался достаточно скудный набор данных, из-за чего модели и выводы могут существенно отличаться при новом, более обширном анализе. С учетом этого, большинство шагов автоматизировано, что позволяет пересчитывать параметры моделей без внесения изменений в код. Для построения моделей и анализа данных использовался язык программирования Python и его библиотеки [11].

3.2 Подготовка данных

До перехода к построению моделей, необходимо подготовить имеющиеся данные. Для этого рассмотрим их повнимательнее. Заметим, что некоторые факторы принимают одинаковые значения на всех объектах выборки, и, следовательно, не несут полезной информации. Удалим их.

Также следует учесть, что некоторые показатели могут быть высоко коррелированы, то есть возможно наличие мультиколлинеарности со всеми вытекающими последствиями для модели (подробнее о влиянии мультиколленеарности написано во второй главе). Будем считать, что два фактора тесно коррелированы между собой, если значение парной корреляции между ними больше или равно 0,85 (данное значение можно гибко изменять в коде программы в зависимости от целей исследователя). Удалим из наборов факторов минимально необходимое для устранения тесных корреляций число показателей. При этом следует иметь в виду, что на данный момент речь идёт лишь о коллинеарности внутри различных наборов факторов, но не между ними. Наглядно исходные корреляции изображены на рисунках 5-7.

После проведённого отбора в наборах остались следующие индикаторы:

- Институциональные факторы:

    а) Уровень парламентской конкуренции;

    б) Включённость граждан в избирательный процесс;

    в) Участие граждан в деятельности общественных организаций;

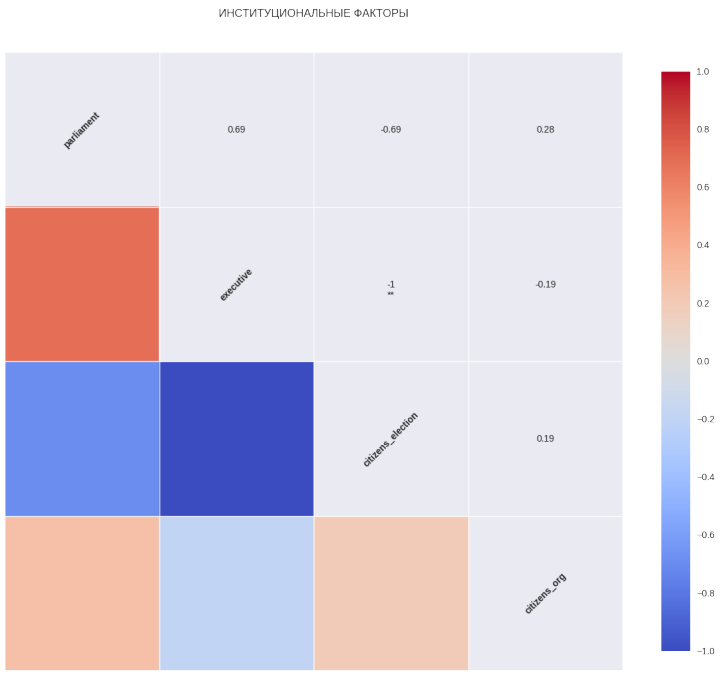

Рисунок 5 — Исходные корреляции между институциональными факторами

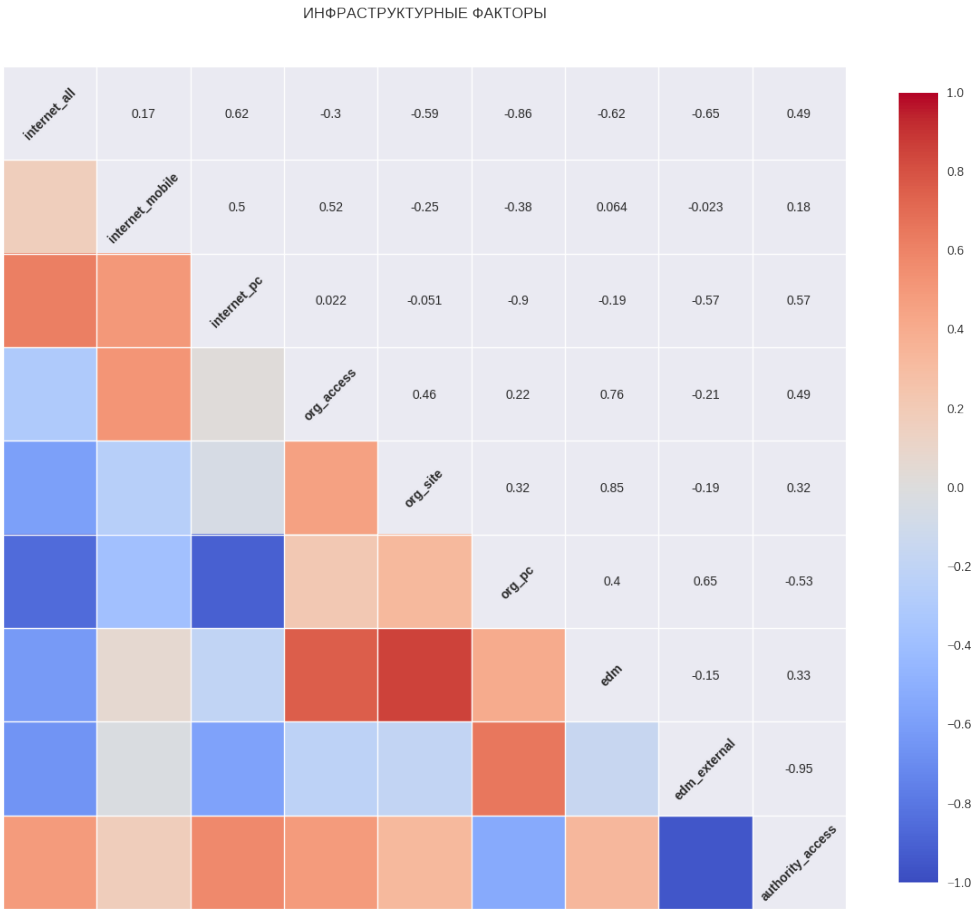

Рисунок 6 — Исходные корреляции между инфраструктурными факторами

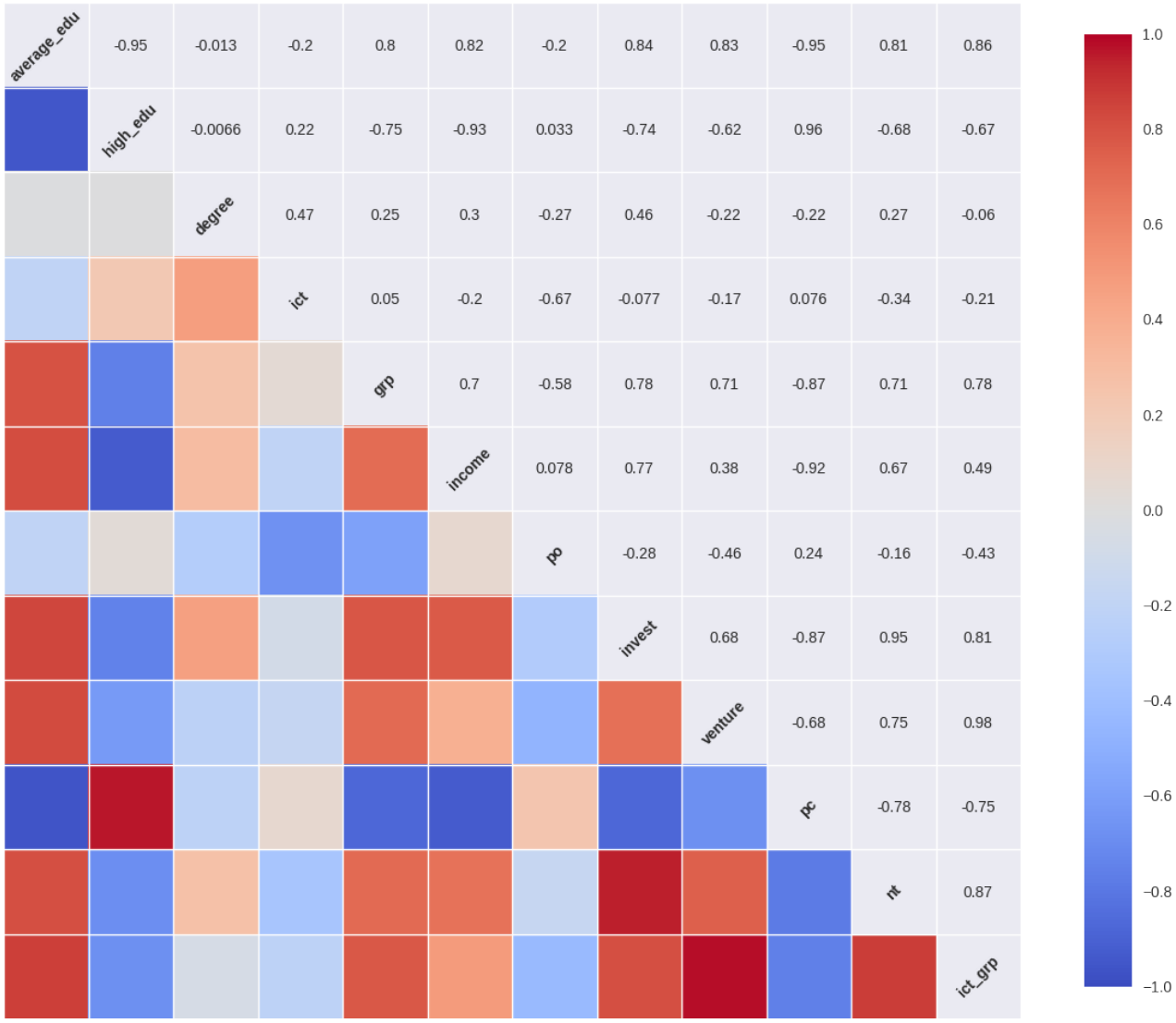

Рисунок 7 — Исходные корреляции между ресурсными факторами

- Инфраструктурные факторы:

а) Количество абонентов сети Интернет;

б) Количество абонентов мобильного интернета;

в) Число персональных компьютеров;

г) Доля организаций, связанных с бизнесом и использующих сеть Интернет;

д) Доля организаций с системой электронного документооборота;

е) Доля организаций с автоматическим обменом данными между своими и внешними информационными системами;

- Ресурсные факторы:

а) Доля населения со средним образованием;

б) Доля населения с учёными степенями;

в) Доля специалистов в области ИКТ;

г) ВРП на душу населения;

д) Среднедушевые доходы населения;

е) Объём использования программного обучения;

ж) Доступность венчурного капитала;

з) Затраты организаций на сетевые технологии.

3.3 Построение моделей для каждого из наборов данных

Теперь для каждого набора по оставшимся факторам строим модели для предсказания значения целевых переменных. В данной работе в качестве классификаторов используются логистическая регрессия и градиентный бустинг, в качестве регрессоров — гребневая регрессия, лассо регрессия и ard регрессия (применялись библиотека sklearn для языка Python и её реализации указанных алгоритмов). За меру качества построенной модели взята средняя среднеквадратическая ошибка и число правильных ответов на скользящем контроле по отдельным объектам. Все эти параметры могут быть легко изменены при запуске программы.

Начнём с построения моделей для непрерывных показателей. Вывод программы для различных наборов факторов и показателей представлен в приложении А.

Изучив полученные результаты, можно сделать следующие выводы:

1) Для показателя числа отменённых конкурсов в общем количестве наилучшей оказалась модель гребневой регрессии по инфраструктурным факторам:

$$cancelled = 0.030 * \text{internet\_all} + 0.946 * \text{internet\_mobile} - 0.505 * \text{internet\_pc} -$$
$$- 0.468 * \text{org\_access} + 0.264 * \text{edm} + 0.201 * \text{edm\_external} \quad ,$$

однако, даже у неё посредственное качество — средняя величина средней квадратичной ошибки достаточно велика;

2) Для показателя числа зарегистрированных организаций ТОС на десять тысяч человек наилучшей оказалась модель ARD регрессии по ресурсным факторам:

$$tsg = 0.262 * \text{average\_edu} + 147.724 * degree - 3.051 * ict + 0.142 * po + 8.957 * venture \quad ,$$

она обладает неплохой предсказывающей силой, так как значение средней средней квадратичной ошибки относительно невелико, и может с успехом применяться на практике для предсказания значения показателя.

Теперь построим модели для бинарного показателя — публичного обсуждения законопроектов в сети Интернет (приложение Б). Наилучшей из них является модель логистической регрессии, построенной по инфраструктурным факторам:

$$\text{public\_discussion} = -0.007 * \text{internet\_all} + 0.459 * \text{internat\_mobile} - 0.168 * \text{internet\_pc} +$$
$$+ 0.067 * \text{org\_access} - 0.347 * edm + 0.063 * \text{edm\_external} \quad ,$$

которая верно угадывает ответ в двух третях случаев, что совсем неплохо, учитывая ограниченный объём анализируемых данных.

3.4 Построение модели для объединённого набора данных

Теперь попробуем построить модели для объединённого набора данных. Ясно, впрочем, что «в лоб» эту задачу решить не удастся, поскольку при объединении наборов число переменных значительно превысит число имеющихся наблюдений, после чего построить вменяемую модель по таким данным вряд ли выйдет. Поэтому,сперва объединим наборы, затем проверим факторы на наличие тесных корреляций между ними и удалим из набора наименьшее их количество для устранения этих корреляций, следом за этим применим метод наименьших квадратов для понижения размерности данных, и лишь затем перейдём к построению моделей.

Корреляции между факторами объединённого набора изображены на рисунке 8.

Рисунок 8 — Корреляции в объединённом наборе данных

После удаления минимально необходимого числа факторов для устранения тесных корреляций в данных, получим новый набор данных, изображённый на рисунке 9.

	parliament	citizens_election	citizens_org	internet_all	internet_mobile	org_access	edm_external	average_edu	degree	ict	grp	income	venture	nt
index														
vo	47.98	63.8	1.67	361.6	58.0	79.2	47.4	65.9	0.015	1.410	278961.2	19055.7	1	2252.6
ra	34.13	64.4	1.01	27.8	70.8	96.2	44.6	63.3	0.016	0.948	174017.6	22054.2	0	46.9
ao	44.02	56.2	0.83	130.2	72.9	86.0	51.0	62.4	0.028	1.320	283591.2	22168.8	0	213.6
rk	36.80	62.0	1.48	24.0	66.6	89.4	51.6	59.2	0.027	1.327	163688.1	12398.4	0	88.8
kk	27.40	70.8	1.10	810.9	82.9	89.2	44.2	68.7	0.031	1.080	330100.2	28787.8	1	12595.5
ro	40.65	63.8	0.96	608.3	57.9	86.0	43.5	63.7	0.035	1.739	235695.9	23354.7	0	657.6

Рисунок 9 — Объединённый набор данных без тесных корреляций

Выходные данные программы после применения метода главных компонент к полученному набору представлен на рисунке 10. Естественно, перед использованием PCA автошкалируем эти данные по причинам, описанным во второй главе.

```
Объединённый набор факторов после стандартизации:

       parliament  citizens_election  citizens_org  internet_all  \
index
vo       1.413273           0.070122      1.655914      0.115474
ra      -0.650751           0.210367     -0.551971     -1.002855
ao       0.823126          -1.706309     -1.154122     -0.659785
rk      -0.252849          -0.350612      1.020310     -1.015586
kk      -1.653703           1.706309     -0.250896      1.620761
ro       0.320905           0.070122     -0.719235      0.941991

       internet_mobile  org_access  edm_external  average_edu    degree  \
index
vo           -1.166028   -1.663581      0.107880     0.692199 -1.394756
ra            0.299618    1.676680     -0.755158    -0.192908 -1.259780
ao            0.540075   -0.327477      1.217500    -0.499291  0.359937
rk           -0.181297    0.340576      1.402436    -1.588654  0.224961
kk            1.685111    0.301278     -0.878449     1.645392  0.764866
ro           -1.177479   -0.327477     -1.094209    -0.056738  1.304772

            ict       grp    income   venture        nt
index
vo     0.422226  0.576730 -0.455805  1.414214 -0.086356
ra    -1.418042 -1.171570  0.152289 -0.707107 -0.574882
ao     0.063732  0.653863  0.175529 -0.707107 -0.537961
rk     0.091615 -1.343654 -1.805899 -0.707107 -0.565602
kk    -0.892251  1.428676  1.517857  1.414214  2.204423
ro     1.732719 -0.144045  0.416029 -0.707107 -0.439622

Координаты главных компонент в пространстве исходных переменных:

[[-0.24456011  0.32555878 -0.07719381  0.3415047   0.23413856  0.01944765
  -0.2679651   0.36820734  0.1033042  -0.11875553  0.28945629  0.3468437
   0.27615964  0.37935759]
 [ 0.38703257 -0.03156947  0.288111    0.20670147 -0.33875586 -0.53973425
   0.02223671  0.15929683 -0.02260044  0.37450739  0.25574479 -0.04654798
   0.28833308  0.00790707]
 [ 0.04875454 -0.19873077 -0.49342     0.20394734 -0.05948883 -0.02407861
  -0.10680413 -0.08402717  0.55173431  0.40733818  0.09844032  0.22148906
  -0.33548506 -0.09408291]]
Доля объясненной дисперсии: [ 0.43967875  0.23325289  0.15586772]
```

Рисунок 10 — Выходные данные программы после применения к объединённому набору факторов метода главных компонент

В результате получим значения первых трёх главных компонент как на рисунке 11.

Теперь перейдём к непосредственной цели нашей работы — моделированию значений показателей по полученным данным. Выходные данные можно увидеть в приложении В.

	0	1	2
index			
vo	-0.318185	3.216722	-1.806857
ra	-0.499549	-2.696540	-1.030017
ao	-1.402636	-0.188467	1.317441
rk	-2.887254	-0.785739	-0.747312
kk	5.126232	-0.553579	-0.263773
ro	-0.018608	1.007602	2.530518

Рисунок 11 — Значения первых трёх главных компонент после применения PCA к объединённому набору данных

Анализируя полученные результаты, можно сделать следующие выводы:

1) На объединённом и обработанном с помощью метода главных компонент наборе данных не удалось построить хорошие модели для контроля над ходом госзакупок и общественного обсуждения законопроектов в сети Интернет;

2) Для количества организаций ТОС на десять тысяч человек была получена модель, построенная на основе гребневой регрессии, отлично описывающая обучающую выборку:

$$tsg = 1.613 * x_0 + 0.609 * x_1 - 1.173 * x_2 \quad .$$

3.5 Экономический анализ полученных результатов

Итак, нами была проделана большая работа по построению и отбору наилучших многомерных моделей регрессии и классификации социально-экономических показателей регионов России. Подведём некоторые итоги выполненной работы.

Во-первых, не удалось построить достаточно качественную модель регрессии доли отменённых конкурсов в общем количестве по имеющимся институциональным, инфраструктурным и ресурсным наборам факторов, а

также по их объединению. Это может свидетельствовать либо об отсутствии зависимости целевого показателя от факторов, либо о неспособности использованных алгоритмов выделить связи и закономерности из-за ограниченного числа наблюдений.

Во-вторых, получилось построить многомерные модели регрессии хорошего качества для числа организаций ТОС на десять тысяч человек как отдельно на наборе ресурсных факторов:

$$tsg = 0.262 * \text{average\_edu} + 147.724 * degree - 3.051 * ict + 0.142 * po + 8.957 * venture \quad ,$$

так и на объединённом и обработанном PCA наборе:

$$tsg = 1.613 * x_0 + 0.609 * x_1 - 1.173 * x_2 \quad .$$

К сожалению, значение главных компонент в данном случае не имеет простой интерпретации, мы можем лишь абстрактно резюмировать, что первая и вторая главные компоненты оказывают положительное влияние на целевую переменную, а третья — отрицательное, и что наибольший вклад в модель вносит первая главная компонента.

Для набора ресурсных факторов всё несколько проще, здесь мы можем лучше объяснить значение полученного результата:

1) При росте доли населения со средним образованием на один процент, число организаций ТОС на десять тысяч человек увеличивается на 0.262 единицы;

2) При росте доли населения, имеющего учёные степени, на один процент, число организаций ТОС на десять тысяч человек увеличивается на 147.724 единицы;

3) При росте доли специалистов, занятых в ИКТ, на один процент, число организаций ТОС на десять тысяч человек уменьшается на три единицы;

4) При росте объёма использования программного обеспечения на один процент, число организаций ТОС на десять тысяч человек увеличивается на 0.142 удиницы;

5) Если венчурный капитал в регионе доступен, то это увеличивает число организаций ТОС на десять тысяч человек на 8.957 единиц;

Наиболее весомый вклад в показатель вносит процент людей с учеными степенями. Вероятно, это объясняется тем, что в силу хорошего образования и высокого уровня интеллектуального развития они значительно лучше других видят плюсы от создания территориального органа самоуправления и соответственно прилагают значительно больше усилий для его формирования. Похожие доводы также применимы для обоснования влияния распространенности среднего образования. Доступность венчурного капитала скорее всего показывает повышенный уровень деловой и финансовой активности в регионе в целом, что положительно сказывается на любой организации, а объем использования программного обеспечения косвенно указывает на доступность сети интернет и наличие доступа к информации о возможности создания ТОС и его преимуществах. Отрицательный вклад процента людей, занятых в ИКТ, можно связать с асоциальностью таких специалистов, недостатком свободного времени из-за большой нагрузки или относительно высоким уровнем заработной платы.

В-третьих, удалось построить модель среднего качества по числу угаданных ответов для индикатора общественного обсуждения законопроектов в сети Интернет:

$$\text{public\_discussion} = -0.007 * \text{internet\_all} + 0.459 * \text{internet\_mobile} - 0.168 * \text{internet\_pc} +$$

$$+0.067 * \text{org\_access} - 0.347 * edm + 0.063 * \text{edm\_external} \quad .$$

Полученный результат говорит о следующем:

1) Наибольшее влияние на вероятность причисления объекта к положительному классу оказывает такой показатель, как количество абонентов мобильного интернета на 100 человек, в несколько раз слабее влияет процент организаций в регионе, связанных с бизнесом, и использующих интернет, а также процент организаций с налаженным обменом информацией между своими и внешними информационными системами;

2) Сильнее всего шанс классификации моделью объекта как отрицательного увеличивает процент организаций с системами электронного документооборота, также существенное влияние на это оказывает число

персональных компьютеров на 100 человек, существует небольшая корреляция с числом абонентов в сети Интернет в тысячах человек.

Интерпретация полученных результатов вызывает некоторые затруднения. Факторы, увеличивающие вероятность классификации объекта как положительного, вопросов не вызывают, поскольку вполне логично, что увеличение числа пользователей мобильного интернета, как и увеличение числа организаций, использующих Интернет, благотворно скажется на публичном обсуждении законопроектов в сети Интернет. Неожиданными являются отрицательные корреляции с такими показателями, как процент организаций с системами электронного документооборота, число персональных компьютеров на 100 человек и число абонентов в сети Интернет в тысячах человек. Эти, на первый взгляд, выпадающие из области разумного факты, скорее всего связаны со способом определения значения целевого показателя в источниках: вероятно, на его значение влияет некоторое соотношение количественных показателей обсуждения в интернете и предполагаемого общего числа пользователей сети, которое на исследованной выборке уменьшается при увеличении числа пользователей.

Заключение

В данной работе был изучен ряд социально-экономических показателей, с использованием современных методов машинного обучения были построены модели многомерной регрессии и классификации этих показателей на наборах институциональных, инфраструктурных и ресурсных факторов, а также на объединённом наборе факторов, предварительно обработанном методом главных компонент с целью уменьшения размерности данных. Полученные модели могут с успехом использоваться для прогнозирования значения рассмотренных показателей или их корректировки посредством изменения значения оказывающих на них влияние факторов.

По итогам проведённого исследования данных, были построены следующие модели:

1) Для числа организаций территориального общественного самоуправления:

а) По набору ресурсных факторов:

$$tsg = 0.262 * \text{average\_edu} + 147.724 * degree - 3.051 * ict + 0.142 * po + 8.957 * venture \ ;$$

б) По объединённому набору факторов после его обработки методом главных компонент для уменьшения размерности данных:

$$tsg = 1.613 * x_0 + 0.609 * x_1 - 1.173 * x_2 \ ;$$

2) Для индикатора общественного обсуждения законопроектов в сети Интернет:

$$\text{public\_discussion} = -0.007 * \text{internet\_all} + 0.459 * \text{internet\_mobile} - 0.168 * \text{internet\_pc} +$$
$$+ 0.067 * \text{org\_access} - 0.347 * edm + 0.063 * \text{edm\_external} \ .$$

К сожалению, не удалось построить достаточно качественную модель регрессии доли отменённых конкурсов в общем количестве по имеющимся институциональным, инфраструктурным и ресурсным наборам факторов, а также по их объединению. Это может свидетельствовать либо об отсутствии зависимости целевого показателя от факторов, либо о неспособности использованных алгоритмов выделить связи и закономерности из-за ограниченного числа наблюдений.

Полученные результаты Результаты работы в первую очередь могут пригодиться органам государственного управления в регионах России, поскольку позволяют оценивать неизвестные значения целевых показателей по известным наборам факторов и учитывать их при планировании работы в регионе.

В целом по итогам моделирования можно сделать следующие выводы:

1) На современном этапе развития методов машинного обучения даже небольшого количества данных достаточно для построения модели, удовлетворительно предсказывающей значение целевого показателя;

2) Лучше всего для построения моделей на небольших наборах данных (размером до тысяч или десятков тысяч точек) подходит регрессия с автоматическим определением значимости признаков (ARD регрессия), но в силу ее большой вычислительной сложности [38] на наборах средних размеров практичнее использовать линейную регрессию с регуляризацией $L1$ (lasso регрессию) или линейную регрессию с $L2$ -регуляризацией (ridge регрессию).

Вследствие подробного изложения использованных методов и пошагового описания принципов и алгоритма работы программы в третьей главе, данная работа может служить практическим руководством по построению многомерных моделей регрессии и классификации социально-экономических показателей.

В процессе выполнения работы поставленные цели и задачи были выполнены в полном объёме.

Список использованных источников

1. Айвазян, С. А. Прикладная статистика и основы эконометрики. Том 1 / С. А. Айвазян.— М.: ЮНИТИ-ДАТА, 2001. — 656 с.;

2. Ардашева, Е. П. Особенности оценки инвестиционного и инновационного развития отрасли в системе управленческого мезоэкономического анализа / Е. П. Ардашева // Вестник Казанского технологического университета. 2007. №3-4. URL: http://cyberleninka.ru/article/n/osobennosti-otsenki-investitsionnogo-i-innovatsionnogo-razvitiya-otrasli-v-sisteme-upravlencheskogo-mezoekonomicheskogo-analiza;

3. Бирюков, А. Н. Нечеткая регрессионная прогнозная многофакторная модель для решения экономической прикладной задачи / А. Н. Бирюков // УЭкС. 2010. №22. URL: http://cyberleninka.ru/article/n/nechetkaya-regressionnaya-prognoznaya-mnogofaktornaya-model-dlya-resheniya-ekonomicheskoy-prikladnoy-zadachi;

4. Богданова, Т. К. Прогнозирование вероятности банкротства предприятий с учетом изменения финансовых показателей в динамике / Т. К. Богданова, Ю. А. Алексеева // Бизнес-информатика. 2011. №1 (15). URL: http://cyberleninka.ru/article/n/prognozirovanie-veroyatnosti-bankrotstva-predpriyatiy-s-uchetom-izmeneniya-finansovyh-pokazateley-v-dinamike;

5. Воронцов, К. В. Курс лекций / К. В. Воронцов [Электронный ресурс]. — Режим доступа: http://www.chemometrics.ru/materials/textbooks/pca.htm, свободный. — Загл. с экрана;

6. Гаврилина, Е. А. Количественная оценка метакомпетенций учащихся на основе методов машинного обучения / Е. А. Гаврилина, М. А. Захаров, А. П. Карпенко // Наука и образование: научное издание МГТУ им. Н. Э. Баумана. 2015. №4. URL:

http://cyberleninka.ru/article/n/kolichestvennaya-otsenka-metakompetentsiy-uchaschihsya-na-osnove-metodov-mashinnogo-obucheniya;

7. Гневашева, В. А. Тенденции трудовых ожиданий и занятости на рынке труда России / В. А. Гневашева // Финансовая аналитика: проблемы и решения. 2013. №41. URL: http://cyberleninka.ru/article/n/tendentsii-trudovyh-ozhidaniy-i-zanyatosti-na-rynke-truda-rossii;

8. Гончарова, И. А. Система сравнительной оценки уровня инновационного состояния промышленного предприятия / И. А. Гончарова // Вестник ВГУИТ. 2015. №3 (65). URL: http://cyberleninka.ru/article/n/sistema-sravnitelnoy-otsenki-urovnya-innovatsionnogo-sostoyaniya-promyshlennogo-predpriyatiya;

9. Гузикова, Л. А. Анализ взаимосвязи инновационной деятельности и уровня информатизации в регионах России / Л. А. Гузикова, А. С. Пантелеев // Общество. Среда. Развитие (Terra Humana). 2015. №4 (37). URL: http://cyberleninka.ru/article/n/analiz-vzaimosvyazi-innovatsionnoy-deyatelnosti-i-urovnya-informatizatsii-v-regionah-rossii;

10. Декина, М. П. Статистический анализ факторов дифференциации оплаты труда в Российской Федерации / М. П. Декина // Известия Санкт-Петербургского государственного экономического университета. 2016. №1 (97). URL: http://cyberleninka.ru/article/n/statisticheskiy-analiz-faktorov-differentsiatsii-oplaty-truda-v-rossiyskoy-federatsii;

11. Документация по библиотеке scikit-learn для машинного обучения с Python [Электронный ресурс]. — Режим доступа: http://scikit-learn.org/stable/, свободный;

12. Иванова, В. О. Корреляционно-регрессионный анализ влияния государственных закупок на инновационное развитие / В. О. Иванова // Проблемы современной экономики. 2013. №3 (47). URL: http://cyberleninka.ru/article/n/korrelyatsionno-regressionnyy-analiz-vliyaniya-gosudarstvennyh-zakupok-na-innovatsionnoe-razvitie;

13. Калинин, А. П. Построение модели зависимости общественно-экономических показателей регионов России от ресурсных факторов / А. П. Калинин // Символ науки. 2016. №12-1. URL: http://os-russia.com/SBORNIKI/SN-16-12-1.pdf;

14. Китова, О. В. Метод машин опорных векторов для прогнозирования показателей инвестиций / О. В. Китова, И. Б. Колмаков, И. А. Пеньков // Статистика и экономика. 2016. №4. URL: http://cyberleninka.ru/article/n/metod-mashin-opornyh-vektorov-dlya-prognozirovaniya-pokazateley-investitsiy;

15. Коложвари, Э. С. Развитие территориального общественного самоуправления как инструмент укрепления социального партнерства власти и населения / Э. С. Коложвари, К. А. Глазычев // Символ науки. 2016. №5-1 (17). URL: http://cyberleninka.ru/article/n/razvitie-territorialnogo-obschestvennogo-samoupravleniya-kak-instrument-ukrepleniya-sotsialnogo-partnerstva-vlasti-i-naseleniya

16. Кузнецов, С. Г. Методология прогнозирования эффективной структуры занятости в условиях инновационного сценария макроэкономического развития / С. Г. Кузнецов, И. И. Мухина // Научные труды: Институт народнохозяйственного прогнозирования РАН. 2008. №6. URL: http://cyberleninka.ru/article/n/metodologiya-prognozirovaniya-effektivnoy-struktury-zanyatosti-v-usloviyah-innovatsionnogo-stsenariya-makroekonomicheskogo;

17. Ляхов, В. П. Территориальное общественное самоуправление в регионах России в условиях реформирования системы местного самоуправления: достижения и риски имитации / В. П. Ляхов // Власть. 2015. №2. URL: http://cyberleninka.ru/article/n/territorialnoe-obschestvennoe-samoupravlenie-v-regionah-rossii-v-usloviyah-reformirovaniya-sistemy-mestnogo-samoupravleniya;

18. Максимов, В. П. Оценка эффективности территориального общественного самоуправления и его вклада в социально-экономическое развитие муниципального образования / В. П. Максимов // Вестник ЧелГУ.

2004. №1. URL: http://cyberleninka.ru/article/n/otsenka-effektivnosti-territorialnogo-obschestvennogo-samoupravleniya-i-ego-vklada-v-sotsialno-ekonomicheskoe-razvitie;

19. Максимов, В. П. Методические вопросы оценки социальной эффективности территориального общественного самоуправления в муниципальном образовании / В. П. Максимов // Вестник ЧелГУ. 2009. №42. URL: http://cyberleninka.ru/article/n/metodicheskie-voprosy-otsenki-sotsialnoy-effektivnosti-territorialnogo-obschestvennogo-samoupravleniya-v-munitsipalnom-obrazovanii;

20. Машинное обучение [Электронный ресурс] // Википедия. — 2017. // — Режим доступа: https://ru.wikipedia.org/wiki/Машинное_обучение, свободный. — Загл. с экрана;

21. Медведева, Н. Г. К вопросу о теории и практике государственного контроля за госзакупками / Н. Г. Медведева // Символ науки. 2015. №7-1. URL: http://cyberleninka.ru/article/n/k-voprosu-o-teorii-i-praktike-gosudarstvennogo-kontrolya-za-goszakupkami;

22. Носко, В. П. Эконометрика. Книга 1. / В. П. Носко. — М.: Издательский дом «Дело» РАНХиГС, 2011. — 672 с.;

23. Орешков, В. И. Совершенствование процесса принятия управленческих решений в экономике и бизнесе на основе применения интеллектуального анализа данных / В. И. Орешков, Е. П. Васильев // Фундаментальные исследования. 2012. №9-4. URL: http://cyberleninka.ru/article/n/sovershenstvovanie-protsessa-prinyatiya-upravlencheskih-resheniy-v-ekonomike-i-biznese-na-osnove-primeneniya-intellektualnogo-analiza;

24. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М.: ЭКЗАМЕН, 2004. — 656 с.;

25. Померанцев, А. Метод главных компонент (PCA) / А. Померанцев [Электронный ресурс] // Российское хемометрическое общество. — 2008. // —

Режим доступа: http://www.chemometrics.ru/materials/textbooks/pca.htm, свободный. — Загл. с экрана;

26. Осколкова, М. А. Возможности прогнозирования динамики фондового индекса s&p 500 с помощью нейросетевых и регрессионных моделей / М. А. Осколкова, П. А. Паршаков // Программные продукты и системы. 2012. №4. URL: http://cyberleninka.ru/article/n/vozmozhnosti-prognozirovaniya-dinamiki-fondovogo-indeksa-s-p-500-s-pomoschyu-neyrosetevyh-i-regressionnyh-modeley;

27. Першин, Д. А. Эволюция системы госзакупок в Российской Федерации / Д. А. Першин // Социально-экономические явления и процессы. 2014. №3. URL: http://cyberleninka.ru/article/n/evolyutsiya-sistemy-goszakupok-v-rossiyskoy-federatsii;

28. Резчиков, А. Ф. Методы прогнозной оценки социально-экономических показателей национальной безопасности / А. Ф. Резчиков, А. Д. Цвиркун, В. А. Кушников, Н. В. Яндыбаева, В. А. Иващенко // Проблемы управления. 2015. №5. URL: http://cyberleninka.ru/article/n/metody-prognoznoy-otsenki-sotsialno-ekonomicheskih-pokazateley-natsionalnoy-bezopasnosti;

29. Рябенко, Е. А. Математика и Python для анализа данных / Е. А. Рябенко, Е. А. Соколов, В. В. Кантор, Э. Драль [Электронный ресурс]. — Режим доступа: https://www.coursera.org/learn/mathematics-and-python/;

30. Рябенко, Е. А. Обучение на размеченных данных / Е. А. Рябенко, Е. А. Соколов, В. В. Кантор, Э. Драль, К. В. Воронцов [Электронный ресурс]. — Режим доступа: https://www.coursera.org/learn/supervised-learning/home/;

31. Токарева, Е. А. Формирование инвестиционного портфеля на основе ценных бумаг РФ / Е. А. Токарева // Символ науки. 2016. №12-1. URL: http://os-russia.com/SBORNIKI/SN-16-12-1.pdf;

32. Троелсен, Е. Язык программирования C# 5.0 и платформа .NET 4.5 / Е. Троелсен. — М.: И. Д. Вильямс, 2013. — 1312 с.;

33. Федотов, В. Х. Нечеткое управление региональными экономическими системами / В. Х. Федотов // Вестник ЧГУ. 2011. №4. URL:

http://cyberleninka.ru/article/n/nechetkoe-upravlenie-regionalnymi-ekonomicheskimi-sistemami;

34. Фридман, Ю. А. Оценка влияния уровня инновационности развития сырьевой отрасли на конкурентоспособность региона / Ю. А. Фридман, Г. Н. Речко, А. Г. Пимонов // Вестник КузГТУ. 2015. №5 (111). URL: http://cyberleninka.ru/article/n/otsenka-vliyaniya-urovnya-innovatsionnosti-razvitiya-cyrievoy-otrasli-na-konkurentosposobnost-regiona;

35. Шрамко, О. Г. Факторный анализ инвестиционной привлекательности регионов / О. Г. Шрамко // Региональная экономика: теория и практика. 2014. №4. URL: http://cyberleninka.ru/article/n/faktornyy-analiz-investitsionnoy-privlekatelnosti-regionov;

36. Яковлев, А. А. Исследования государственных закупок в России / А. А. Яковлев // Экономический журнал ВШЭ. 2014. №4. URL: http://cyberleninka.ru/article/n/issledovaniya-gosudarstvennyh-zakupok-v-rossii;

37. Jordan, M. I. Machine learning: Trends, perspectives, and prospects / M. I. Jordan, T. M. Mitchell [Electronic resource] // Science, vol. 349. — 2015. // — Access: https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf, free. — Title from screen;

Machine Learning [Electronic resource] // Wikipedia. — 2017. // — Access: https://en.wikipedia.org/wiki/Machine_learning, free. — Title from screen;

38. MacKay, D. Bayesian Methods for Adaptive Models / D. MacKay [Electronic resource] // California Institute of Technology. — 1992. // — Access: http://www.inference.phy.cam.ac.uk/mackay/thesis.pdf, free. — Title from screen;

39. Mitchell, T. M. Machhine Learning / T. M. Mitchell // McGraw Hill. — 1997. // — 432 с.;

40. Mohri, M. Foundations of Machine Learning / M. Mohri, A. Rostamizadeh, A. Talwalkar // MIT Press, 2012. — 412 с.;

41. Munoz, A. Machine Learning and Optimization / A. Munoz [Electronic resource]. — Access: https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf, free. — Title from screen;

42. Ng, A. Machine Learning by Stanford University / A. Ng [Electronic resource]. — Access: https://www.coursera.org/learn/machine-learning/home/welcome, free. — Title from screen;

43. Risvik, H. Principal Component Analysis (PCA) & NIPALS algorithm / H. Risvik [Electronic resource] // 2007. // — Access: http://folk.uio.no/henninri/pca_module/pca_nipals.pdf, free. — Title from screen;

44. Samuel, A. L. Some studies in machine learning using the game of checkers / A. L. Samuel // IBM Journal of research and development. — 1959.

45. Shlens, J. A Tutorial on Principal Component Analysis / J. Shlens [Electronic resource] // Center for Neural Science, New York University and Systems Neurobiology Laboratory, Salk Insitute for Biological Studies. — 2009. // — Access: http://www.snl.salk.edu/~shlens/pca.pdf, free. — Title from screen;

46. Smith, L. A tutorial on Principal Component Analysis / L. A. Smith [Electronic resource] // University of Otago. — 2002. // Access: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, free. — Title from screen;

47. Souza, C. A Tutorial on Principal Component Analysis with the Accord.NET Framework / C. Souza [Electronic resource] // Federal University of Sao Carlos, Center of Science and Technology, Computer Department. — 2012 // Access: http://www2.dc.ufscar.br//~cesar.souza/publications/pca-tutorial.pdf, free. — Title from screen;

48. Valiant, L. G. A theory of the learnable / L. G. Valiant // Communications of the ACM. — 1984. — pp. 1134-1142;

49. Vapnik, V. N. Statistical learning theory / V. N. Vapnik. — N.Y.: John Wiley & Sons, Inc., 1998;

50. Wilhelm, F. Explaining the idea behind automatic relevance determination and bayesian interpolation / F. Wilhelm [Electronic resource] // PyData Amsterdam. —2016 // Access: http://www.slideshare.net/FlorianWilhelm2/explaining-the-idea-behind-automatic-

relevance-determination-and-bayesian-interpolation-59498957, free. — Title from the screen.

ПРИЛОЖЕНИЕ А

Выходные данные программы при моделировании непрерывных показателей по наборам факторов

```
Набор данных: Инфраструктурные факторы.
Прогнозируемая переменная: cancelled.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -10.47797098   -5.91386179  -68.88680089  -56.47205841 -614.78754184
  -31.72066524]

Средняя величина среднеквадратической ошибки:

-131.37648319

Коэффициенты:

[ 0.02383206  0.8174205  -0.38438579 -0.07941918  0.02386181  0.        ]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -22.36244846   -9.0470868   -96.23159258  -25.24453496 -505.12793941
  -53.71051063]

Средняя величина среднеквадратической ошибки:

-118.620685473

Коэффициенты:

[ 0.02972234  0.9462321  -0.50491445 -0.46849838  0.26419339  0.20143681]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -1.46281777e+01  -1.96182019e-01  -1.69072006e+02  -2.83128291e+01
  -5.34409253e+02  -2.49804935e+01]

Средняя величина среднеквадратической ошибки:

-128.599823443

Коэффициенты:

[  2.33738529e-02   7.97771641e-01  -3.68125809e-01  -1.51429749e-03
   0.00000000e+00   1.55595373e-04]
```

Рисунок А.1 — построение модели для контроля над ходом госзакупок по инфраструктурным факторам.

```
Набор данных: Институциональные факторы.
Прогнозируемая переменная: cancelled.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -26.71168736 -131.94217971 -389.11387229  -13.43656837 -600.20950347
  -13.68295311]

Средняя величина среднеквадратической ошибки:

-195.849460718

Коэффициенты:

[-0.58710839  0.85818828 -0.        ]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -38.33888715 -135.10319896 -446.13823109  -14.72159927 -644.30490668
  -19.42366109]

Средняя величина среднеквадратической ошибки:

-216.33841404

Коэффициенты:

[-0.55818963  0.94634389 -0.81428373]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -52.22363008 -131.20187156 -464.72895913  -18.26102326 -613.54743935
  -18.17656678]

Средняя величина среднеквадратической ошибки:

-216.356581693

Коэффициенты:

[ -4.68264060e-01   7.36115226e-01  -4.13597796e-04]
```

Рисунок А.2 — построение модели для контроля над ходом госзакупок по институциональным факторам.

```
Набор данных: Ресурсные факторы.
Прогнозируемая переменная: cancelled.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -3.25755947e+01  -8.59130491e+01  -4.41890769e+01  -1.48627951e+02
  -1.12152352e+03  -5.04032575e-03]

Средняя величина среднеквадратической ошибки:

-238.805706023

Коэффициенты:

[ -7.09863710e-01   0.00000000e+00  -0.00000000e+00  -8.79736414e-06
   2.07502742e-04   0.00000000e+00  -0.00000000e+00   2.26779175e-03]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[  -32.44008293   -59.50898623   -47.98829055  -155.29278743 -1231.84707672
    -1.73169575]

Средняя величина среднеквадратической ошибки:

-254.801486604

Коэффициенты:

[ -1.18973713e+00   5.33135839e-03   1.92050251e-02   1.74494562e-06
   2.82036062e-04   2.57167301e-02  -2.40000988e-01   2.39117651e-03]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -678.37457879 -1903.67968031    -4.16614787   -48.65516257  -676.47650878
  -192.45598233]

Средняя величина среднеквадратической ошибки:

-583.968010108

Коэффициенты:

[ -2.12128073e-02   1.04903817e+03  -1.97740997e+01   0.00000000e+00
   0.00000000e+00   2.32248174e-01   1.49551029e+01   0.00000000e+00]
```

Рисунок А.3 — построение модели для контроля над ходом госзакупок по ресурсным факторам.

```
Набор данных: Инфраструктурные факторы.
Прогнозируемая переменная: tsg.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -80.9530893  -164.13197414 -124.67253204  -18.20046175 -229.61446145
 -128.28228672]

Средняя величина среднеквадратической ошибки:

-124.309134233

Коэффициенты:

[ 0.01625921  0.46871645 -0.19945193 -0.93796939  0.42715787  0.        ]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-100.58823605 -243.31507113 -206.00009737  -28.02331484 -671.52247903
 -134.28428368]

Средняя величина среднеквадратической ошибки:

-230.622247017

Коэффициенты:

[ 0.02125552  0.59819368 -0.32150827 -1.301896    0.63907524  0.14137744]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-128.79821611 -217.91956743  -76.8209985   -22.75369148 -275.26638074
 -144.24872235]

Средняя величина среднеквадратической ошибки:

-144.301262767

Коэффициенты:

[ 0.          0.05833868  0.         -0.00150938 -0.00221252 -0.10732998]
```

Рисунок А.4 — построение модели для числа зарегистрированных организаций ТОС по инфраструктурным факторам.

```
Набор данных: Институциональные факторы.
Прогнозируемая переменная: tsg.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-102.96290626   -8.23308617  -83.8872805   -11.12868069 -144.19992241
  -22.67409487]

Средняя величина среднеквадратической ошибки:

-62.1809951499

Коэффициенты:

[ 0.11075081  0.91168285  0.        ]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-122.17673      -6.91389395 -120.65002782  -13.14545375 -139.21880893
  -29.96894957]

Средняя величина среднеквадратической ошибки:

-72.0123106702

Коэффициенты:

[ 0.19249314  1.04636696 -0.06352907]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-196.20987003   -9.13599538 -195.84793307  -16.16539699 -145.01372463
  -16.98709715]

Средняя величина среднеквадратической ошибки:

-96.5600028738

Коэффициенты:

[  6.30024281e-04   7.32598463e-01   7.58861908e-04]
```

Рисунок А.5 — построение модели для числа зарегистрированных организаций ТОС по институциональным факторам.

```
Набор данных: Ресурсные факторы.
Прогнозируемая переменная: tsg.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[  -36.02047651  -100.51360346   -50.4087605   -182.68561104 -1245.30653225
    -2.69277898]

Средняя величина среднеквадратической ошибки:

-269.604627127

Коэффициенты:

[  7.56443958e-01  -0.00000000e+00  -0.00000000e+00   9.19001468e-06
  -3.01889221e-04   1.16955079e-02   0.00000000e+00   6.82179660e-04]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[  -36.48761891   -88.29329939   -46.19878958  -183.31814378 -1306.51565178
    -9.38151701]

Средняя величина среднеквадратической ошибки:

-278.36583674

Коэффициенты:

[  1.25490052e+00  -6.10248890e-03  -4.15477629e-02   2.44377370e-05
  -5.91015746e-04   2.07296731e-01   2.55939335e-01   5.03799089e-04]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-20.08770517 -40.48740111  -0.63639422 -23.82801113 -29.8509674
  -6.54906823]

Средняя величина среднеквадратической ошибки:

-20.2399245414

Коэффициенты:

[  2.62013040e-01   1.47724083e+02  -3.05147598e+00   0.00000000e+00
   0.00000000e+00   1.41682719e-01   8.95740846e+00   0.00000000e+00]
```

Рисунок А.6 — построение модели для числа зарегистрированных организаций ТОС по ресурсным факторам.

ПРИЛОЖЕНИЕ Б

Выходные данные программы при моделировании бинарных показателей по наборам факторов

```
Набор данных: Инфраструктурные факторы.
Прогнозируемая переменная: public_discussion.

LOGREG

Число правильных ответов для каждого тренировочного/тестового набора:

[ 1.  0.  1.  1.  0.  1.]

Среднее число правильных ответов:

0.666666666667

Коэффициенты:

[-0.58710839  0.85818828 -0.        ]


Gradient Boosting

Число правильных ответов для каждого тренировочного/тестового набора:

[ 1.  0.  0.  0.  0.  1.]

Среднее число правильных ответов:

0.333333333333

Коэффициенты:

[-0.55818963  0.94634389 -0.81428373]
```

Рисунок Б.1 — построение модели для общественного обсуждения законопроектов в сети Интернет по инфраструктурным факторам.

```
Набор данных: Институциональные факторы.
Прогнозируемая переменная: public_discussion.

LOGREG

Число правильных ответов для каждого тренировочного/тестового набора:

[ 0.  0.  0.  0.  0.  0.]

Среднее число правильных ответов:

0.0

Коэффициенты:

[-0.58710839  0.85818828 -0.        ]


Gradient Boosting

Число правильных ответов для каждого тренировочного/тестового набора:

[ 0.  1.  0.  0.  0.  1.]

Среднее число правильных ответов:

0.333333333333

Коэффициенты:

[-0.55818963  0.94634389 -0.81428373]
```

Рисунок Б.2 — построение модели для общественного обсуждения законопроектов в сети Интернет по институциональным факторам.

```
Набор данных: Ресурсные факторы.
Прогнозируемая переменная: public_discussion.

LOGREG

Число правильных ответов для каждого тренировочного/тестового набора:

[ 0.  0.  0.  0.  0.  1.]

Среднее число правильных ответов:

0.166666666667

Коэффициенты:

[-0.58710839  0.85818828 -0.        ]


Gradient Boosting

Число правильных ответов для каждого тренировочного/тестового набора:

[ 0.  0.  0.  0.  0.  0.]

Среднее число правильных ответов:

0.0

Коэффициенты:

[-0.55818963  0.94634389 -0.81428373]
```

Рисунок Б.3 — построение модели для общественного обсуждения законопроектов в сети Интернет по ресурсным факторам.

ПРИЛОЖЕНИЕ В

Выходные данные программы при моделировании на объединённых и обработанных методом главных компонент наборах факторов

```
Набор данных: Объединённый набор данных после применения PCA.
Прогнозируемая переменная: cancelled.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -91.91639463 -151.95808412   -1.26903332 -100.14358779 -586.21125924
  -34.34406314]

Средняя величина среднеквадратической ошибки:

-160.973737042

Коэффициенты:

[ 3.11745868 -0.33879925 -0.        ]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-457.8098332  -195.16969016   -3.49655281 -103.09192104 -788.02885375
  -57.95874629]

Средняя величина среднеквадратической ошибки:

-267.592599542

Коэффициенты:

[ 3.19344898 -0.61370499 -0.28970643]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -18.24724355 -225.64409461   -1.1031312  -109.26429212 -839.49894046
  -21.5212917 ]

Средняя величина среднеквадратической ошибки:

-202.546498941

Коэффициенты:

[  3.12437931e+00  -1.23294831e-03  -2.52706247e-04]
```

Рисунок В.1 — построение модели для контроля над ходом госзакупок по объединённому и обработанному PCA набору факторов

```
Набор данных: Объединённый набор данных после применения PCA.
Прогнозируемая переменная: tsg.

LASSO

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -1.95901149e+01  -6.43670063e+00  -2.55076985e-03  -3.61665007e-03
  -8.90110224e+01  -2.08011228e+01]

Средняя величина среднеквадратической ошибки:

-22.6408546825

Коэффициенты:

[ 1.49380485  0.33381952 -0.80386663]


RIDGE

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[ -1.41199378  -2.27534953  -0.80368654  -0.12246496 -19.14800458
  -6.20604851]

Средняя величина среднеквадратической ошибки:

-4.9945913147

Коэффициенты:

[ 1.61259834  0.60896707 -1.17257269]


ARD

Значения среднеквадратической ошибки для каждого тренировочного/тестового набора:

[-13.0207251   -3.18195537  -1.31796059  -0.18942619 -10.45636085
  -4.77959327]

Средняя величина среднеквадратической ошибки:

-5.49100356313

Коэффициенты:

[ 1.64799831  0.59973907 -1.23154364]
```

Рисунок В.2 — построение модели для количества организаций ТОС по объединённому и обработанному PCA набору факторов

```
Набор данных: Объединённый набор данных после применения PCA.
Прогнозируемая переменная: public_discussion.

LOGREG

Число правильных ответов для каждого тренировочного/тестового набора:

[ 0.  0.  0.  0.  0.  0.]

Среднее число правильных ответов:

0.0

Коэффициенты:

[[ 0.08392927 -0.27345128  0.07974312]]


Gradient Boosting

Число правильных ответов для каждого тренировочного/тестового набора:

[ 1.  0.  1.  0.  0.  0.]

Среднее число правильных ответов:

0.333333333333
```

Рисунок В.3 — построение модели для общественного обсуждения законопроектов в сети Интернет по объединённому и обработанному PCA набору факторов

ПРИЛОЖЕНИЕ Г

Результативность исследований

1) Отдельные результаты выпускной квалификационной работы (магистерской диссертации) отражены в публикациях:

а) Калинин, А. П. Оценка развития электронного правительства в России / А. П. Калинин, А. В. Сычева. — Региональная научно-практическая конференция «Корпоративная российская модель управления: эффективность, кризисы, риски», ВПИ (ф) ВолгГТУ, 2015 г.;

б) Калинин, А. П. Автоматизация процесса построения моделей зависимости показателей на примере общественной активности в регионах Российской Федерации / А. П. Калинин, А. А. Полковников. — «Инфраструктурное обеспечение социально-экономического развития региона», ВГИ (ф) ВолГУ, 2016 г.;

в) Калинин, А. П. Построение модели зависимости общественно-экономических показателей регионов России от ресурсных факторов / А. П. Калинин // Символ науки. 2016. №12-1. URL: http://os-russia.com/SBORNIKI/SN-16-12-1.pdf;

2) Отдельные результаты магистерской диссертации докладывались на конференциях:

а) Региональная научно-практическая конференция «Корпоративная российская модель управления: эффективность, кризисы, риски», Волжский, 2015 г.;

б) Международная научно-практическая конференция «Инфраструктурное обеспечение социально-экономического развития региона», Волжский, 2016 г.

Руководитель работы _____ доцент Полковников А.А.