Оглавление

Введение……………………………………………………………………...6

1 Аспекты многомерной классификации………………………...9

1.1 Территориальное общественное самоуправление……………………..9

1.2 Государственные закупки………………………...……………………..9

2 Анализ алгоритмов и методов машинного обучения…………………..26

2.1 Основные понятия и обозначения……………………………………..26

2.2 Линейный модели……………………………………………………….

2.2.1 Модель линейной регрессии………………………………………….

2.2.2 Линейный классификатор…………………………………………….

2.3 Проблема переобучения………………………………………………..

2.4 Метрики качества в машинном обучении…………………………….

2.5 Метод максимизации правдоподобия…………………………………..

2.6 Метод автоматического отбора значимых признаков…………………

2.7 Метод главных компонент………………………………………………

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

3.1 Описание исходных социально-экономических данных………………

3.2 Подготовка данных……………………………………………………….

3.3 Построение моделей для каждого из наборов данных…………………

3.4 Построение модели для объединённого набора данных……………….

3.5 Экономический анализ полученных результатов……………………….

Заключение……………………………………………………………………

Список использованных источников………………………………………..

Приложение А. Выходные данные программы при моделировании непрерывных показателей по наборам факторов………………………………

Приложение Б. Выходные данные программы при моделировании бинарных показателей по наборам факторов………………………….

Приложение В. Выходные данные программы при моделировании на объединённых и обработанных методом главных компонент наборах факторов…………………………………………………………………….……….

Приложение Г. Результативность исследований…………………………..

Приложение Д. Справка о проверке выпускной квалификационной работы на предмет наличия заимствований……………………………………..

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 Аспекты многомерной классификации

1.1 Социально-экономические показатели и их роль

Каждое экономическое явление, каждый процесс чаще определяется не одним, обособленным, а целым комплексом взаимосвязанных показателей. Например, эффективность использования основных средств производства характеризует уровень фондоотдачи, фондоёмкости, рентабельности, производительности труда и т.д. В связи с этим выбор и обоснование системы показателей для отражения экономических явлений и процессов (объектов исследования) является важным методологическим вопросом в экономическом анализе. От того, насколько показатели полно и точно отражают сущность изучаемых явлений, зависят результаты анализа.

Так как в анализе используется большое количество разнокачественных показателей, необходима их группировка и систематизация.

По своему содержанию показатели делятся на количественные и качественные. К количественным показателям относятся, например, объём изготовленной продукции, количество работников, площадь посевов, поголовье скота и т. д. Качественные показатели определяют существенные особенности и свойства изучаемых объектов. Примером качественных показателей являются производительность труда, себестоимость, рентабельность, урожайность культур и др.

Изменение количественных показателей обязательно приводит к изменению качественных, и наоборот. Так, например, рост объёма производства продукции ведёт к снижению себестоимости. Рост производительности труда обеспечивает увеличение объёма производства продукции.

Одни показатели используются при анализе деятельности всех отраслей национальной экономики, другие — только в отдельных отраслях. По этому признаку они делятся на общие и специфические. К общим относятся показатели валовой продукции, производительности труда, прибыли, себестоимости и др. Примером специфических показателей для отдельных отраслей и предприятий могут быть калорийность каменного угля, влажность торфа, жирность молока, урожайность культур и т. д.

Показатели, используемые в экономическом анализе, по степени синтеза делятся также на обобщающие, частные и вспомогательные (косвенные). Первые применяются для обобщённой характеристики сложных экономических явлений. Частные показатели отражают отдельные стороны, элементы изучаемых явлений и процессов. Например, обобщающими показателями производительности труда являются среднегодовая, среднедневная, часовая выработка продукции одним рабочим. К частным показателям производительности труда относятся затраты рабочего времени на производство единицы продукции отдельного вида или количество произведённой продукции за единицу рабочего времени.

Вспомогательные (косвенные) показатели используются для более полной характеристики того или иного объекта анализа.

Например, количество рабочего времени, затраченного на единицу выполненных работ.

Аналитические показатели делятся на абсолютные и относительные. Абсолютные показатели выражаются в денежных, натуральных измерителях или через трудоёмкость. Относительные показатели дают соотношение каких-либо двух абсолютных показателей. Определяются в процентах, коэффициентах или индексах

Абсолютные показатели делятся, в свою очередь, на натуральные и стоимостные. Натуральные показатели выражают величину явления в физических единицах измерения (масса, длина, объём и т. д.). Стоимостные показатели характеризуют величину сложных явлений в денежном измерении.

При изучении причинно-следственных отношений показатели бывают факторными и результативными.

Если тот или иной показатель рассматривается как результат воздействия одной или нескольких причин и выступает в качестве объекта исследования, то при изучении взаимосвязей он называется результативным.

Показатели, которые определяют поведение результативного показателя и выступают в качестве причин изменения его величины, называются факторными.

По способу формирования различают показатели нормативные (нормы расхода сырья, материалов, топлива, энергии, нормы амортизации, цены и др.); плановые (данные планов экономического и социального развития предприятия, хозрасчетные задания внутрихозяйственным подразделениям); учётные (данные бухгалтерского, статистического, оперативного учёта); отчётные (данные бухгалтерской, статистической и оперативной отчётности); аналитические (оценочные), которые исчисляются в ходе самого анализа для оценки результатов и эффективности работы предприятий.

Все показатели, которые используются в анализе, взаимосвязаны и взаимообусловлены. Это вытекает из реально существующих связей между экономическими явлениями, которые они описывают. Комплексное изучение экономики предприятий предусматривает систематизацию показателей, потому что совокупность показателей, какой бы исчерпывающей она ни была, без учёта их взаимосвязи, соподчиненности не может дать настоящего представления об эффективности хозяйственной деятельности. Необходимо, чтобы конкретные данные о разных видах деятельности были органически увязаны между собой в единой комплексной системе.

Все показатели в зависимости от объекта анализа группируются в блоки и подсистемы.

Показатели, которые образуют подсистемы, можно разбить на входящие и выходящие, общие и частные. С помощью входящих и выходящих показателей осуществляется взаимосвязь подсистем.

Выходящий показатель одной системы является входящим для других подсистем.

Так, например, показатели первой подсистемы характеризуют производственную структуру предприятия, структуру управления, уровень концентрации и специализации производства, продолжительность производственного цикла, техническую и энергетическую вооружённость труда, степень механизации и автоматизации, прогрессивность технологических процессов и т. д.

В современных условиях, кроме того, надо учитывать природно-климатические условия функционирования объекта.

Показатели приведённой подсистемы оказывают влияние на все остальные показатели хозяйствования и в первую очередь на объём производства и реализацию продукции, её качество, на степень использования производственных ресурсов (производительность труда, фондоотдачу, материалоотдачу), а также на другие показатели экономической эффективности: себестоимость, прибыль, рентабельность и т. д. Поэтому анализ хозяйственной деятельности надо начинать с изучения этой подсистемы.

Показатели второй подсистемы характеризуют обеспеченность предприятия основными средствами производства, оборудованием, материалами и трудовыми ресурсами. От того, насколько полно и своевременно обеспечивается производство всеми необходимыми средствами, зависит выход продукции, себестоимость, прибыль, рентабельность, финансовое положение предприятия.

Третья подсистема включает показатели производства продукции: объём валовой и товарной продукции в стоимостном, натуральном и условно- натуральном измерении, структуру продукции, её качество, ритмичность производства и др.

К показателям четвёртой подсистемы относятся данные об объёме отгрузки и реализации продукции, остатках готовой продукции на складах. Они очень тесно связаны с показателями третьей и девятой подсистем.

Основными показателями пятой подсистемы являются фондоотдача, фондоёмкость, среднегодовая стоимость основных средств производства, амортизация. Наряду с этими показателями большое значение имеют и другие, например, среднечасовая выработка продукции на единицу оборудования, коэффициент использования наличного оборудования и т. д. От них зависит уровень фондоотдачи.

В шестой подсистеме основными показателями являются материалоёмкость, материалоотдача, стоимость использованных предметов труда за анализируемый отрезок времени. Они тесно связаны с показателями подсистем 7, 8, 9, 10. От экономического использования материалов зависит выход продукции, себестоимость, а значит и сумма прибыли, уровень рентабельности, финансовое состояние предприятия.

К показателям седьмой подсистемы относятся показатели использования трудовых ресурсов.

Основными показателями восьмой подсистемы считаются показатели себестоимости продукции.

В девятой подсистеме основными показателями являются показатели прибыли и рентабельности.

К последней подсистеме (блок 10) относятся показатели, которые характеризуют наличие и структуру капитала предприятия по составу его источников и формам размещения, эффективность и интенсивность использования собственных и заёмных средств. Отношения с финансовыми органами, банками, вышестоящими организациями, работниками своего предприятия по поводу использования собственного и заёмного капитала, а также факторы, которые определяют эти отношения.

К этой подсистеме относятся также показатели, которые характеризуют использование прибыли, фондов накопления и потребления, кредитов банков, платёжеспособность предприятия, его финансовую устойчивость. Они зависят от показателей всех предыдущих подсистем и в свою очередь оказывают большое влияние на показатели организационно-технического уровня предприятия, объём производства продукции, эффективность использования материальных и трудовых ресурсов.

Таким образом, все показатели хозяйственной деятельности предприятия находятся в тесной связи и зависимости, которую необходимо учитывать в комплексном анализе. Взаимосвязь основных показателей определяет последовательность выполнения анализа от изучения первичных показателей до обобщающих. Такая последовательность соответствует объективной основе формирования экономических показателей.

Например, чтобы определить плановый объём производства продукции, надо знать условия и возможности производства, его обеспеченность необходимыми средствами производства в требуемых пропорциях и достигнутый уровень использования средств и предметов труда. Только с учётом этого можно точно обосновать объём производства продукции. Себестоимость единицы продукции можно исчислять, зная затраты труда, материалов, сырья, сумму амортизации и др., а также объём производства продукции. Финансовый результат можно определить после реализации продукции путём сравнения суммы выручки с суммой затрат на производство и реализацию продукции и т.д. В такой последовательности формируются показатели при составлении плана социального и экономического развития предприятия, в такой последовательности должен проводиться и анализ хозяйственной деятельности. Но это не исключает и обратную последовательность анализа — от обобщающих показателей к частным. Главное, чтобы при этом была обеспечена системность, учитывалась взаимосвязь отдельных блоков анализа между собой и достигалось единство результатов анализа по каждому разделу.

В экономических исследованиях под фактором обычно понимают движущие силы какого-либо процесса, явления, определяющие его характер или отдельные черты.

Оценить происходящие процессы, дать им оценку можно через показатели. Поэтому иначе это можно определить так: факторы – частные показатели, характеризующие условия, причины изменения результативного показателя.

Чтобы глубже выяснить воздействие факторов на хозяйственную деятельность объектов анализа, нужно обеспечить их научно обоснованную классификацию и систематизацию. В экономической литературе принята следующая их группировка:

1. По степени воздействия на результаты работы объектов факторы подразделяются на основные и второстепенные.

2. По степени количественного измерения влияния факторов на совокупный результативный показатель они делятся на поддающиеся и не поддающиеся количественной оценке.

3. По времени воздействия факторы бывают постоянные, влияющие непрерывно на весь изучаемый период, и временные, действующие на определенный период.

4. По охвату действием различаются общие и специфические факторы.

5. По степени детализации они делятся на простые, являющиеся результатом действия одной причины, и сложные, возникающие из-за комплекса причин.

6. По характеру действия факторы делятся на объективные, не зависящие от результатов работы объекта, и субъективные, зависящие от них.

7. По способу определения подразделяются на прямые, влияние которых изучается непосредственно без специальных приемов, и расчетные, действие которых измеряется при помощи специальных расчетов.

8. По действию на результаты хозяйственной деятельности факторы могут быть положительными и отрицательными.

9. По степени действия факторы могут быть первого, второго, третьего, n-го порядка.

В отдельную группу выделяют два вида следующих факторов.

10. По направлениям путей развития факторы можно подразделить на интенсивные и экстенсивные.

11. По уровням объекта управления факторы можно подразделить на: технико-экономические; ресурсов; процессов; социально-организа­ционные; социально-психологические; физиологические и др.

По цели измерения влияния на результаты хозяйственной деятельности факторы разделяют на количественные и качественные, сложные и простые, прямые и косвенные, измеримые и неизмеримые.

В экономической литературе и на практике практически разработана методика факторного анализа (табл. 2.2).

Таблица 1 — Описание методов факторного анализа

|  |  |
| --- | --- |
| Метод факторного анализа | Описание |
| 1 | 2 |
| Детерминированный (функциональный) | Представляет собой методику исследования влияния факторов, связь которых с результа­тивным показателем носит функциональный характер, т.е. результативный показатель мо­жет быть представлен в виде произведения, частного или алгебраической суммы факторов |
| Стохастический (корреляционный) | Использование факторов, связь которых явля­ется неполной, вероятностной (корреляцион­ной). Если при функциональной (полной) зависимости с изменением аргумента всегда происходит соответствующее изменение функции, то при корреляционной связи изменение аргумента может дать несколько значений прироста функции в зависимости от сочетания других факторов, определяющих данный показатель. Например, производительность труда при одном уровне фондовооруженности может быть неодинаковой на разных предприятий. Это зависит от оптимальности сочетания других факторов, воздействующих на этот показатель |
| Прямой (дедуктивный) | Исследование ведется дедуктивным способом – от общего к частному |
| Обратный индуктивный | Осуществляет исследование причинно-следст­венных связей способом логичной индукции – от частных, отдельных факторов к обобщаю­щим |
| Одноступенчатый | Используется для исследования факторов толь­ко одного уровня (одной ступени) подчинения без их детализации на составные части. Напри­мер, y=a\*b |
| Многоступенчатый | Проводится детализация факторов a и b на со­ставные элементы с целью изучения их поведе­ния. Детализация факторов может быть продол­жена и дальше. В данном случае изучается вли­яние факторов различных уровней соподчинен­ности |
| Статический | Применяется при изучении влияния факторов на результативные показатели на соответству­ющую дату |
| Динамичный | Представляет собой методику исследования причинно-следственных связей в динамике |
| Ретроспективный | Изучает причины прироста результативных показателей за прошлые периоды |
| Перспективный (прогнозный) | Исследует поведение факторов и результатив­ных показателей в перспективе |
| Основные | Оказывают решающее воздействие на результа­тивный показатель |
| Второстепенные | Не оказывают решающего воздействия на ре­зультаты хозяйственной деятельности в данных условиях. Здесь необходимо заметить, что один и тот же фактор в зависимости от обстоятельств может быть и основным, и второстепенным. |
| Внутренние | Зависят от деятельности предприятия |
| Внешние | Не зависят от деятельности предприятия |
| Объективные | Не зависят от воли и желаний людей, например стихийное бедствие |
| Субъективные | Зависят от деятельности отдельных людей, предприятий, организаций и учреждений |
| Общие | Действуют во всех отраслях экономики |
| Специфические | Действуют в условиях отдельной отрасли эко­номики или предприятия |
| Постоянные | Оказывают влияние на изучаемое явление бес­прерывно на протяжении всего времени |
| Переменные | Воздействие переменных факторов проявляется периодически, например, освоение новой тех­ники, новых видов продукции, новой техноло­гии производства и т. д. |
| Экстенсивные | Связаны с количественным, некачественным приростом результативного показателя, напри­мер, увеличение объема производства продук­ции путем расширения производственных пло­щадей, увеличения станков, количества рабо­чих и т. д. |
| Интенсивные | Характеризуют степень усилия, напряженности труда в процессе производства, например, по­вышение уровня производительности труда |
| Количественные | Выражают количественную определенность явлений (количество рабочих, оборудования, сырья и т. д.) |
| Качественные | Определяют внутренние качества, признаки и особенности изучаемых объектов (производи­тельность труда, качество продукции, плодоро­дие почвы и т. д.) |
| Сложные | Примером сложного фактора является произво­дительность труда |
| Простые | Количество рабочих дней в отчетном периоде |
| Прямые | Непосредственно влияющие на результаты де­ятельности |
| Косвенные | Опосредованно влияющие на результаты дея­тельности |
| Измеряемые | Подверженные количественному измерению |

Относительные показатели отражают соотношение величины изучаемого явления с величиной какого-либо другого явления или с величиной этого явления, но взятой за другое время или по другому объекту. Относительные показатели получают в результате деления одной величины на другую, которая принимается за базу сравнения. Это могут быть данные плана, базисного года, другого предприятия, среднеотраслевые и т.д. Относительные величины выражаются в форме коэффициентов (при базе 1) или процентов (при базе 100).

В анализе хозяйственной деятельности используются разные виды относительных величин: планового задания, выполнения плана, динамики, структуры, координации, интенсивности, эффективности.

Относительная величина планового задания представляет собой отношение планового уровня показателя текущего года к его уровню в прошлом году или к среднему за три-пять предыдущих лет.

Относительная величина выполнения плана – это отношение между фактическим и плановым уровнем показателя, выраженное обычно в процентах.

Для характеристики изменения показателей за какой-либо промежуток времени используют относительные величины динамики. Их определяют путем деления величины показателя текущего периода на его уровень в предыдущем периоде (месяце, квартале, году). Называются они темпами роста (прироста) и выражаются обычно в процентах или коэффициентах. Относительные величины динамики могут быть базисными и цепными. В первом случае каждый следующий уровень динамического ряда сравнивается с базисным годом, а в другом – каждый следующий год относится к предыдущему.

Показатель структуры – это относительная доля (удельный вес) части в общем, выраженная в процентах или коэффициентах. Например, удельный вес активной части основных средств в общей их стоимости, удельный вес рабочих в общем количестве работников предприятия, удельный вес прямых затрат в общей себестоимости продукции и другие показатели.

Относительные величины координации представляют собой соотношение частей целого между собой, например, активной и пассивной части основных производственных фондов, силовых и рабочих машин и т.д.

Относительными величинами интенсивности называются те, которые характеризуют степень распространенности, развития какого-либо явления в соответствующей среде, например, степень заболеваемости населения, процент рабочих высшей квалификации и т.д.

Относительные величины эффективности – это соотношение эффекта с ресурсами или затратами, например, производство продукции (рождаемость, смертность, производительность и др. показатели) на одного работника, жителя, на один рубль затрат, на одного рабочего и т.д.

Разновидностью относительных величин являются индексы.

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.

Объектомиз пространства объектовназывается элемент, для которого требуется сделать предсказание.

Ответомиз пространства допустимых ответовназывается результат предсказания модели на объекте.

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

Обучающей выборкой называется совокупность известных прецедентов, то есть множество объектов и соответствующих им ответов.

Тогда задачей обучения по прецедентам является восстановление зависимости целевой функциипо обучающей выборкепутём построения решающей функции, которая приближала бы не только на объектах обучающей выборки, но и на всём множестве. При этом решающая функциядолжна быть эффективно реализуема на компьютере, вследствие чего некоторые авторы называют её алгоритмом.

Признаковым описанием объекта называется совокупность его признаков (-мерный вектор): , где  — признак, некоторая числовая характеристика, описывающая объект [37, с.. С формальной точки зрения признак — это отображение, где  — множество допустимых значений признака. Далее множество допустимых значений-го признака будем обозначать.

Совокупность признаковых описаний всех объектов выборки, записанную в виде матрицы размера, называют матрицей «объекты-признаки»:

.

Вместе с ней обычно работают с вектором истинных ответов для всех объектов выборки:

.

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

- бинарные признаки,;

- вещественные признаки,;

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

- порядковые признаки, — упорядоченное множество;

- множественные признаки,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 ошибкиназывается некоторая характеристика качества работы алгоритмана выборке .

Тогда задача машинного обучения сводится к подбору такого алгоритмаиз некоторого семействаалгоритмов, для которого достигается минимум функционала ошибки.

Формализованное определение задачи обучения на размеченных данных звучит следующим образом. Дана обучающая выборка, необходимо найти такой алгоритм, на котором будет достигаться минимум функционала ошибки:

.

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

- соответствует задаче бинарной классификации;

- соответствует задаче многоклассовой классификации;

- ,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:

,

где — свободный коэффициент, — признаки, а — их веса.

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

,

где — 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, но функция модуля не является гладкой функцией, что вызывает сложности с оптимизацией градиентными методами. Поэтому мерой ошибки чаще всего выбирают квадрат отклонения:

.

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

.

Который для линейной модели принимает вид:

.

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

.

Среднеквадратичная ошибка может быть записана в матричном виде:

.

Можно найти аналитическое решение задачи минимизации:

.

Однако, тогда придётся вычислять обратную матрицу, что требует порядкаопераций дляпризнаков, 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 итерациииз приближения, полученного на предыдущей итерации, вычитается вектор градиента в соответствующей точке, домноженный на коэффициент, называемый шагом:

.

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

.

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:

.

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

.

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-ое слагаемое для-й компоненты показывает, как надо изменить вес, чтобы улучшить качество для-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:

.

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:

.

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

.

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 объектадо гиперплоскостиравно: .

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

.

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

.

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 оценкупороговой функции и минимизируют не долю неправильных ответов, а некоторую функцию, являющуюся оценкой сверху:

.

Примерами таких функций потерь являются:

- логистическая функция потерь:;

- экспоненциальная функция потерь: ;

- кусочно-линейная функция потерь (используется в методе опорных векторов): .

В случае логистической функции потерь функционал ошибки имеет вид:

.

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, что для любогоиз выборки выполняется:

.

Допустим, было найдено некоторое решение задачи оптимизации:

.

Но тогда для вектор весов получим другое решение задачи оптимизации:. 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 регуляризациик значению функционала ошибки с последующей минимизацией не функционала ошибки, а полученной суммы. Например, для квадратичного регуляризатора, называемого также-регуляризатором, задача оптимизации примет вид:

,

где .

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, достаточно большого, чтобы не допустить переобучения, но при этом не зануляющего значимых коэффициентов и позволяющего обнаружить закономерности в данных.

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

.

Заметим, что, в отличии от-регуляризатора, -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 начастей 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 полученопоказателей 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 регуляризациив 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):

.

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):

.

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 векторемодель сумела объяснить:

,

где .

Для разумных моделей коэффициент детерминации лежит в пределах от нуля до единицы, причёмсоответствует идеальной модели,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):

.

Параметропределяет, за что штрафовать сильнее — за недопрогноз или за перепрогноз. Еслиближе 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):

.

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, где  — 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

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | True Positive (TP) | False Positive (FP) |
|  | 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:

.

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

.

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:

.

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:

.

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-мерой:

.

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

.

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

Большое количество алгоритмов бинарной классификации сперва вычисляют вещественное число, являющееся оценкой вероятности принадлежности классу, а затем сравнивают его с некоторым порогом:

.

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-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-кривых для задач с небольшим и значительным количеством объектов представлены на рисунке Ч.

Заметим, что-кривая выходит всегда из начала координат, а заканчивается точкой, где — 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(area under curve — precision-recall curve).

Другим способом оценки принадлежности к классуявляется-кривая, на оси абсциcc которой откладывается(false positive rate), а на оси ординат —(true positive rate):

, .

Примеры-кривых представлены на рисунке Ч. Заметим, что любая-кривая исходит из точкии приходит в точку, а идеальный классификатор пройдёт через точку. 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 величинаимеет распределение, — выборка размера:

, .

Тогда функция правдоподобия имеет вид:

.

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

.

Тогда оценкой максимального правдоподобия называется величина:

.

Для непрерывной случайной величины метод максимального правдоподобия записывается аналогично:

, , .

Метод максимального правдоподобия обладает следующими свойствами:

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

при .

- асимптотическая нормальность, то есть с ростом объёма выборки оценки максимального правдоподобия всё лучше описываются нормальным распределением со средним, равным истинному значению, и дисперсией, равной величине, обратной к информации Фишера:

при .

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

.

Если этот случайный шум имеет нормальное распределение с нулевым средним и дисперсией, то задача минимизации среднеквадратичной ошибки даёт оценку максимального правдоподобия для регрессионной функции :

.

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 — константа, апредставляет собой случайную функцию с плотностью распределения. Тогда среднеквадратичная ошибка примет вид:

.

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

.

Если же — произвольная функция от признаков, то то функционал среднеквадратичной ошибки примет вид:

,

а его минимум будет представлять собой условное математическое ожидание:

.

Таким образом, в случае конечной выборки:

,

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

.

В случае линейной регрессии, когда отклик моделируется линейной комбинацией, наилучшей линейной аппроксимацией условного математического ожиданияявляется выражение:

, .

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:

.

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:

 — лучшая аппроксимация.

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:

 — лучшая аппроксимацияпорядка.

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

,

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-регуляризатором) соответствует заданию нормального априорного распределения на коэффициенты линейной модели, а метода лассо (lasso, регрессия с-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:

,

где  — априорное распределение, играющее роль регулятора весов.

Действительно, при использовании ARD семейство матриц расширяется:

.

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 — прямоугольной таблицы чисел размерностьюстрок истолбцов. 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:

.

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, c. 217].

Итак, пусть имеется матрица переменныхразмерностью, где — число образцов, а  — количество независимых переменных (столбцов), которых, как правило, много (). В метода главных компонент используются новые формальные переменные(), являющиеся линейной комбинацией исходных переменных():

.

С помощью этих новых переменных матрицаразлагается в произведение двух матрици: .

Матрицаназывается матрицей счетов (scores), матрицаназывается матрицей нагрузок, а матрица — 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

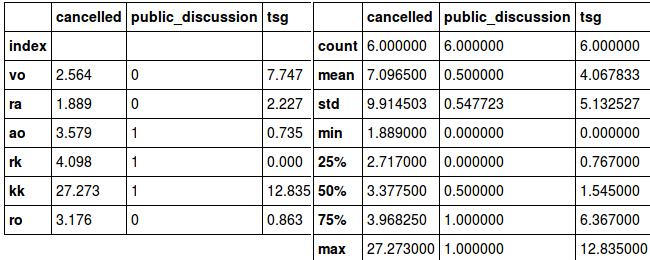
Современные методы машинного обучения позволяют строить модели зависимостей для точного предсказания целевых переменных по набору факторов. Это используется в совершенно разных областях науки и в огромном количестве практических задач. В экономике их применяют преимущественно для оценки значения показателей, 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.

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;

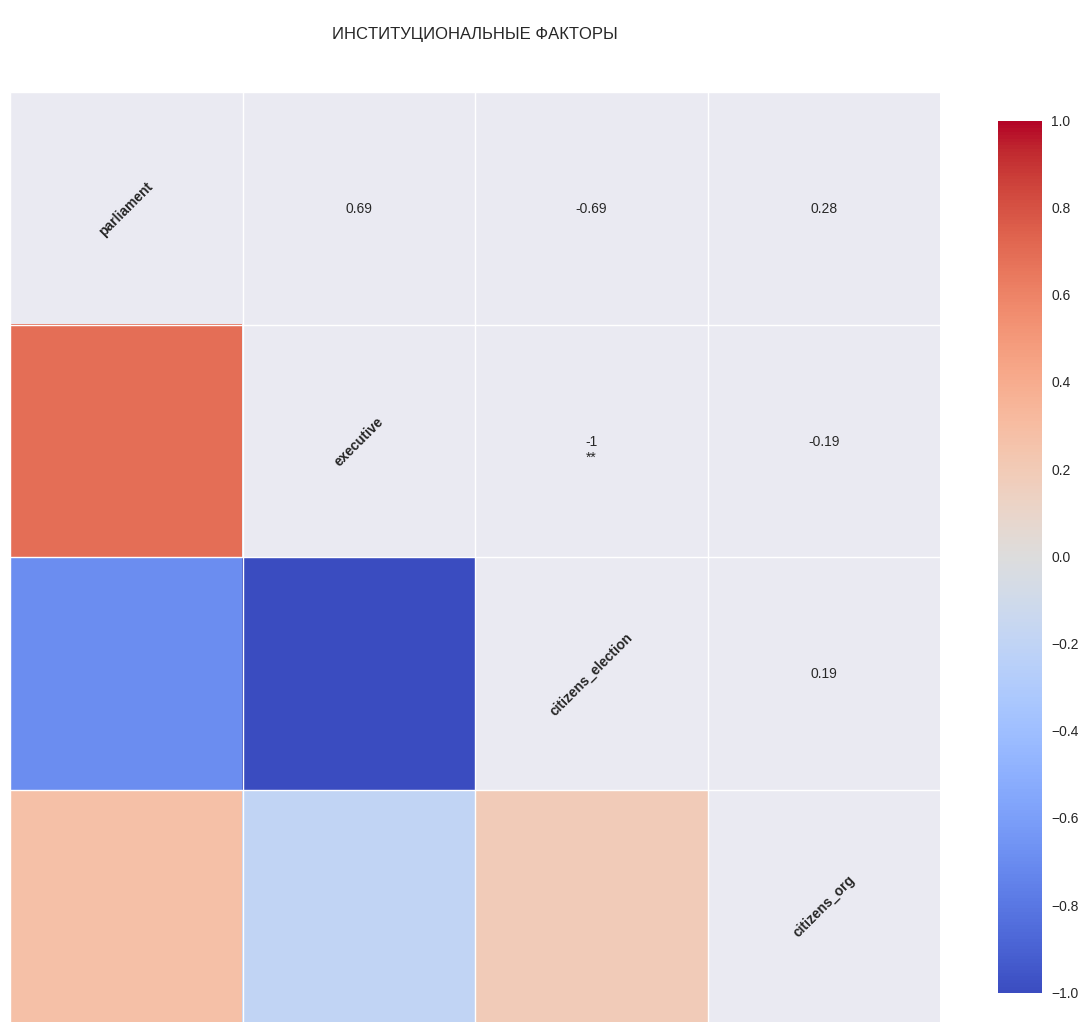


Рисунок 5 — Исходные корреляции между иституциональными факторами

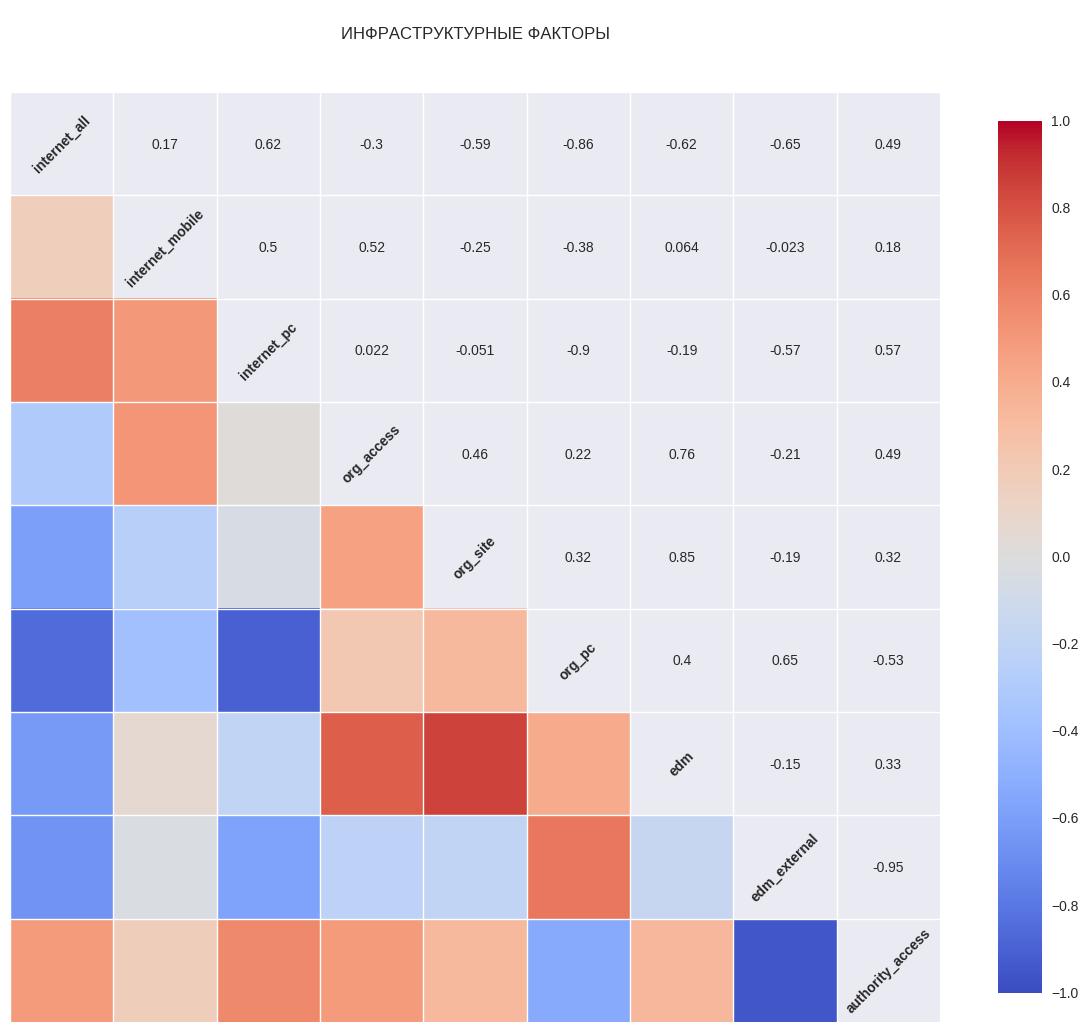
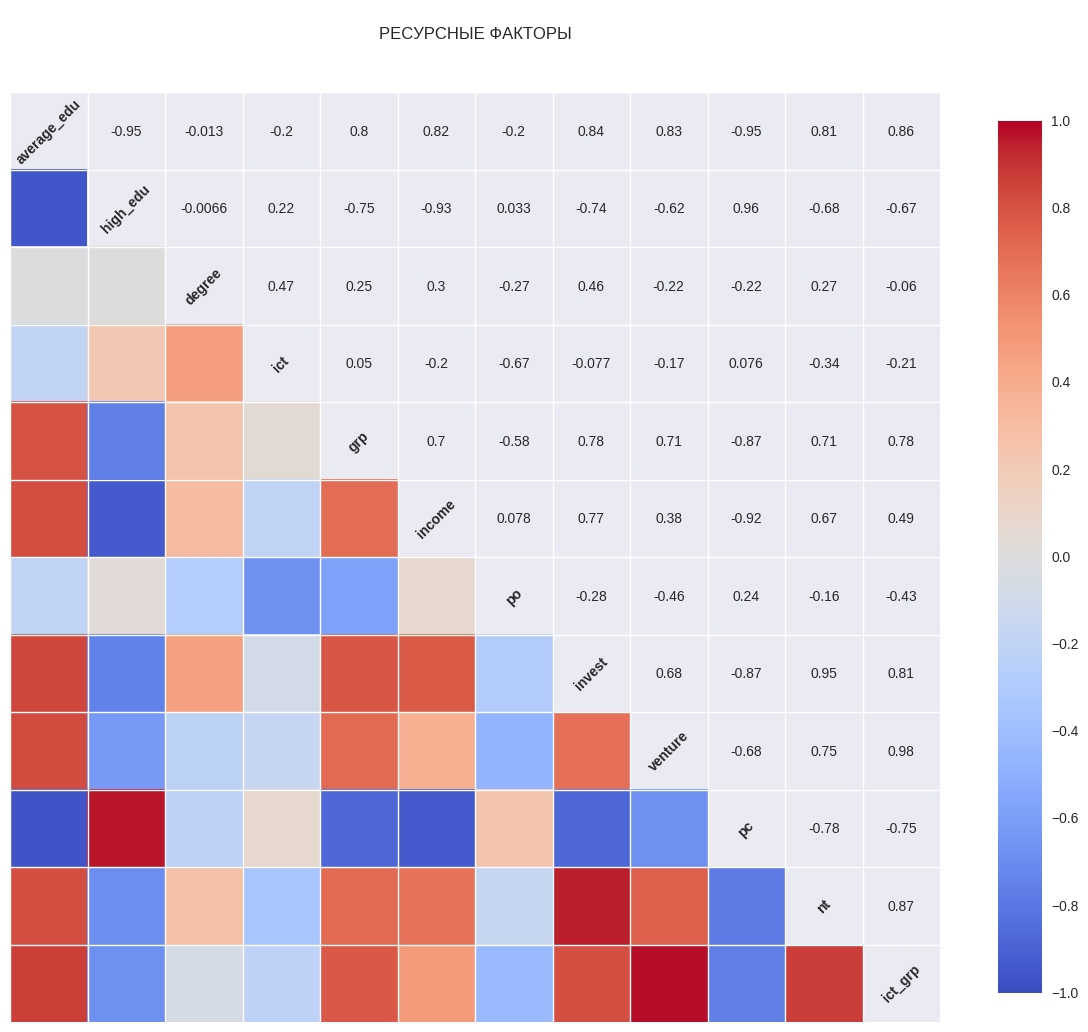


Рисунок 6 — Исходные корреляции между инфраструктурными факторами

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:

,

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:

,

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:

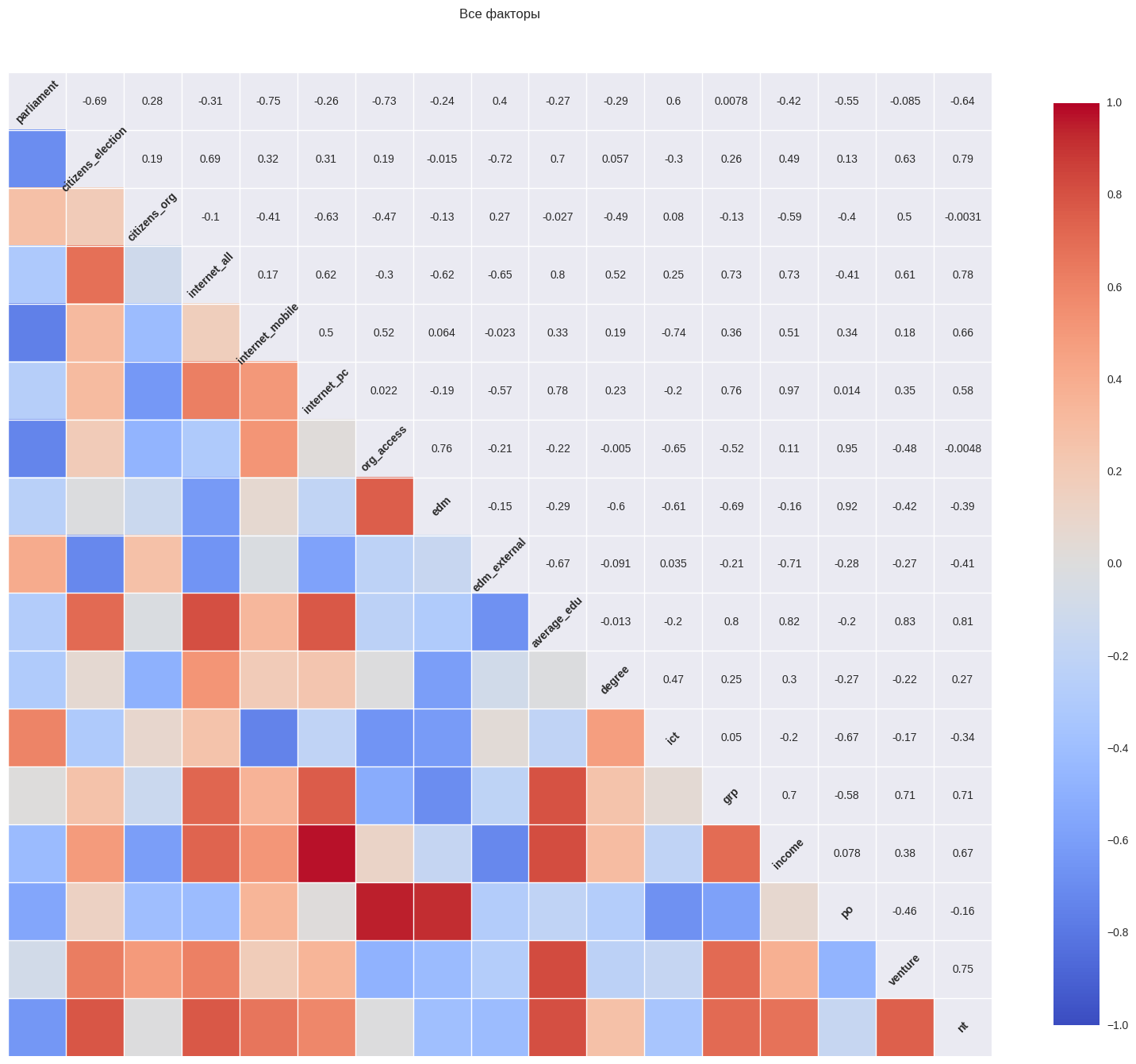
,

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

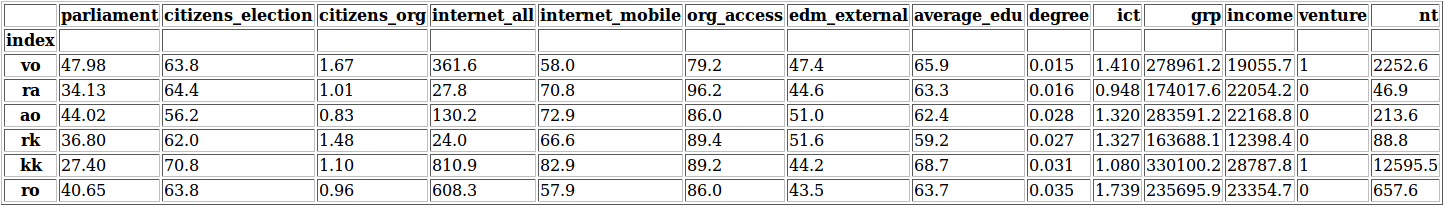
3.4 Построение модели для объединённого набора данных

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.

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.

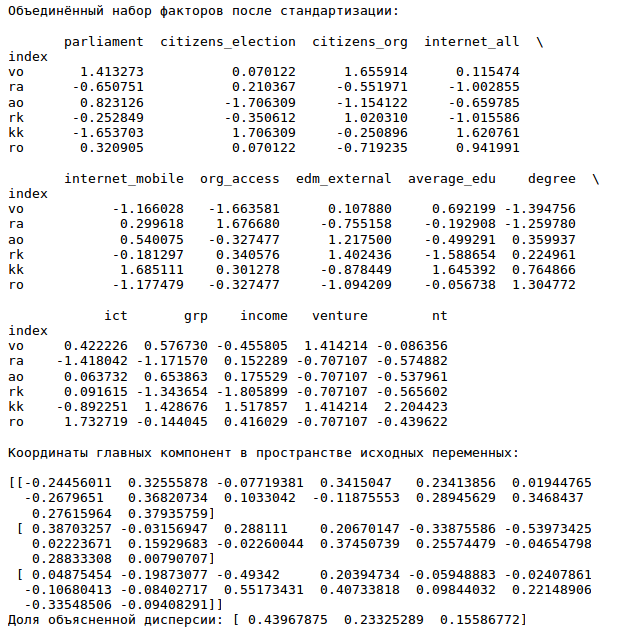


Рисунок 10 — Выходные данные программы после применения к объединённому набору факторов метода главных компонент

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.

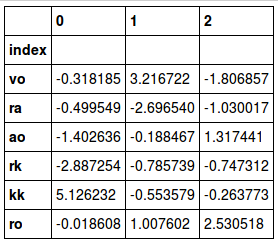


Рисунок 11 — Значения первых трёх главных компонент после применения PCA к объединённому набору данных

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:

.

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:

,

так и на объединённом и обработанном PCA наборе:

.

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:

.

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 иституциональных, 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:

;

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

;

2) Для индикатора общественного обсуждения законопроектов в сети Интернет:

.

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(lasso регрессию) или линейную регрессию с-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;

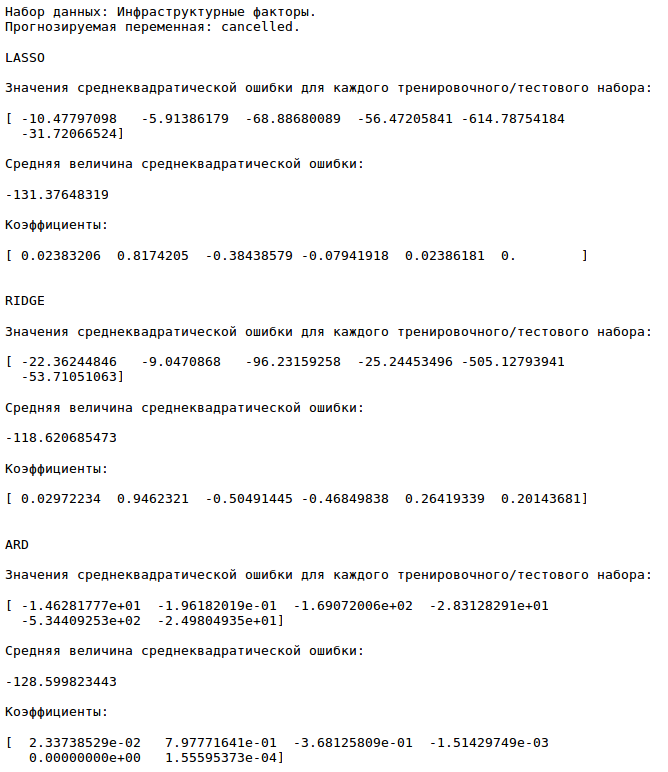
1. Медведева, Н. Г. К вопросу о теории и практике государственного контроля за госзакупками / Н. Г. Медведева // Символ науки. 2015. №7-1. URL: <http://cyberleninka.ru/article/n/k-voprosu-o-teorii-i-praktike-gosudarstvennogo-kontrolya-za-goszakupkami>;
2. Носко, В. П. Эконометрика. Книга 1. / В. П. Носко. — М.: Издательский дом «Дело» РАНХиГС, 2011. — 672 с.;
3. 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>;
4. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М.: ЭКЗАМЕН, 2004. — 656  с.;
5. 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;
6. 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>;
7. 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>;
8. 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>;
9. 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/>;
10. 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/>;
11. 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>;
12. Троелсен, Е. Язык программирования C# 5.0 и платформа .NET 4.5 / Е. Троелсен. — М.: И. Д. Вильямс, 2013. — 1312 с.;
13. 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>;
14. Фридман, Ю. А. Оценка влияния уровня инновационности развития cырьевой 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>;
15. 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>;
16. 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>;
17. 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;

1. 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;
2. Mitchell, T. M. Machhine Learning / T. M. Mitchell // McGraw Hill. — 1997. // — 432 с.;
3. Mohri, M. Foundations of Machine Learning / M. Mohri, A. Rostamizadeh, A. Talwalkar // MIT Press, 2012. — 412 с.;
4. 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;
5. 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;
6. 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;
7. Samuel, A. L. Some studies in machine learning using the game of checkers / A. L. Samuel // IBM Journal of research and development. — 1959.
8. 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;
9. 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;
10. 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;
11. Valiant, L. G. A theory of the learnable / L. G. Valiant // Communications of the ACM. — 1984. — pp. 1134-1142;
12. Vapnik, V. N. Statistical learning theory / V. N. Vapnik. — N.Y.: John Wiley & Sons, Inc., 1998;
13. 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

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

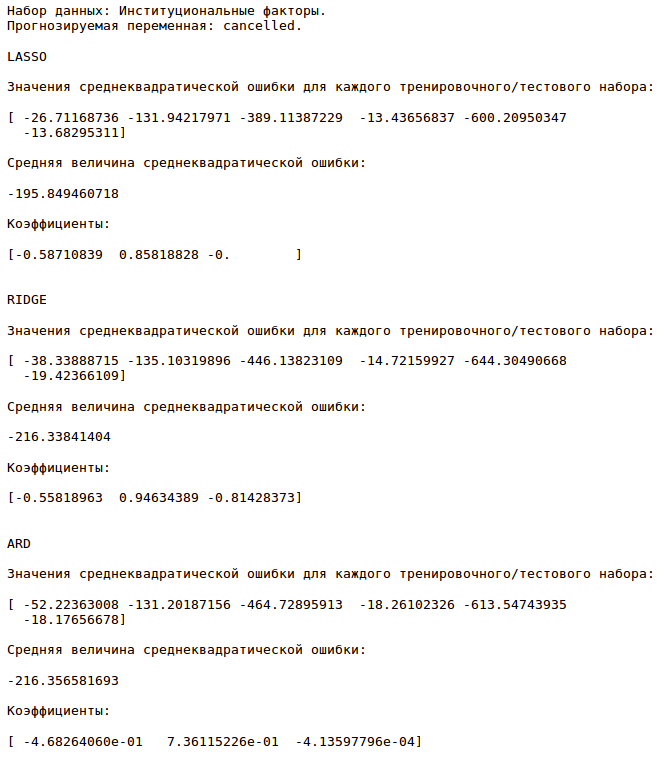


Рисунок А.2 — построение модели для контроля над ходом госзакупок по институциональным факторам.

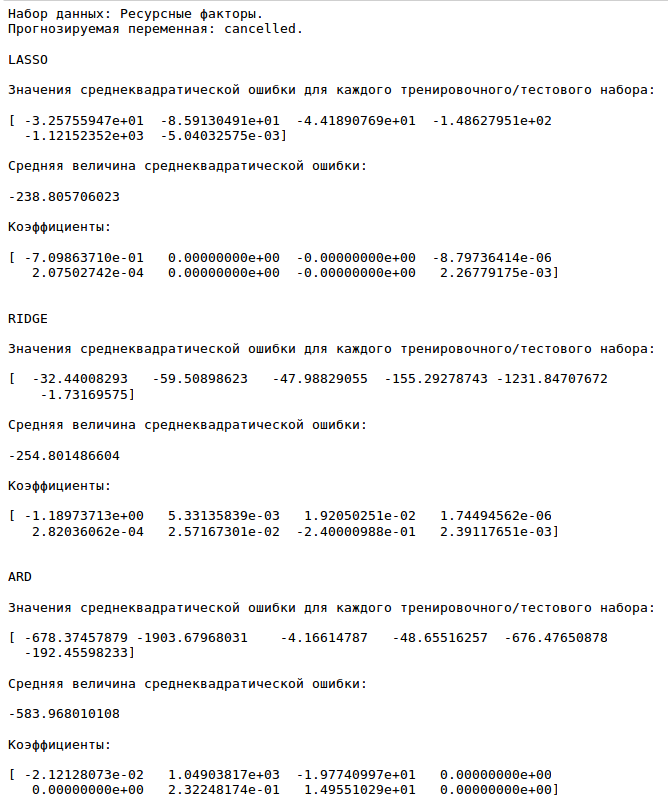


Рисунок А.3 — построение модели для контроля над ходом госзакупок по ресурсным факторам.

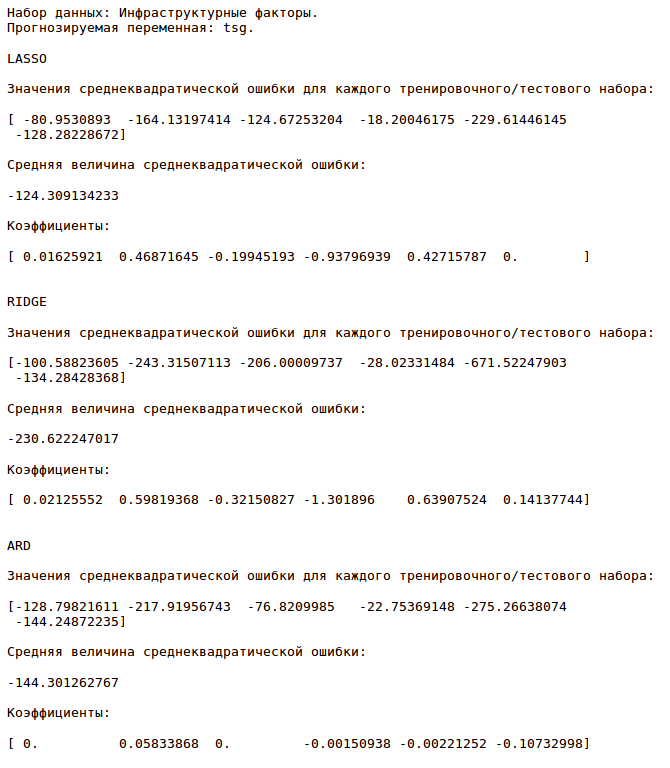


Рисунок А.4 — построение модели для числа зарегистрированных организаций ТОС по инфраструктурным факторам.

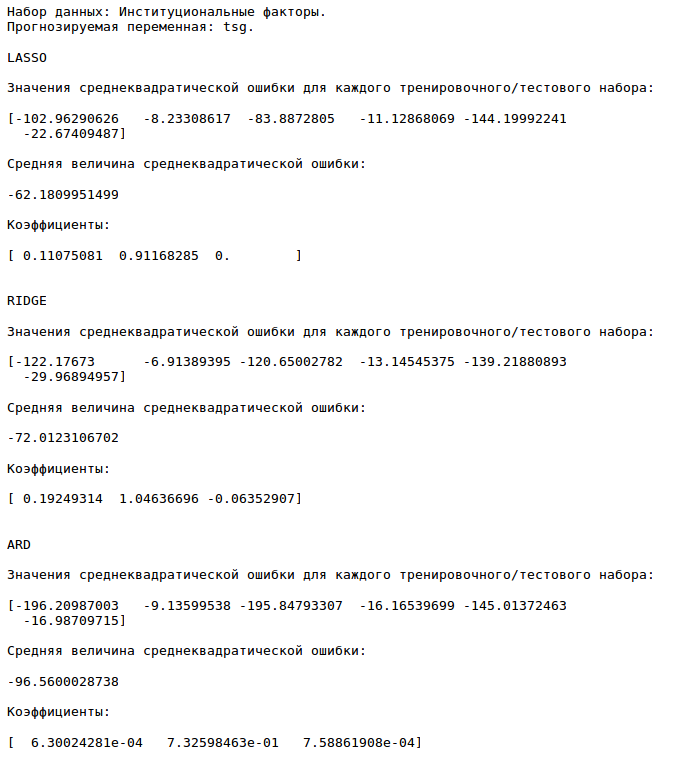
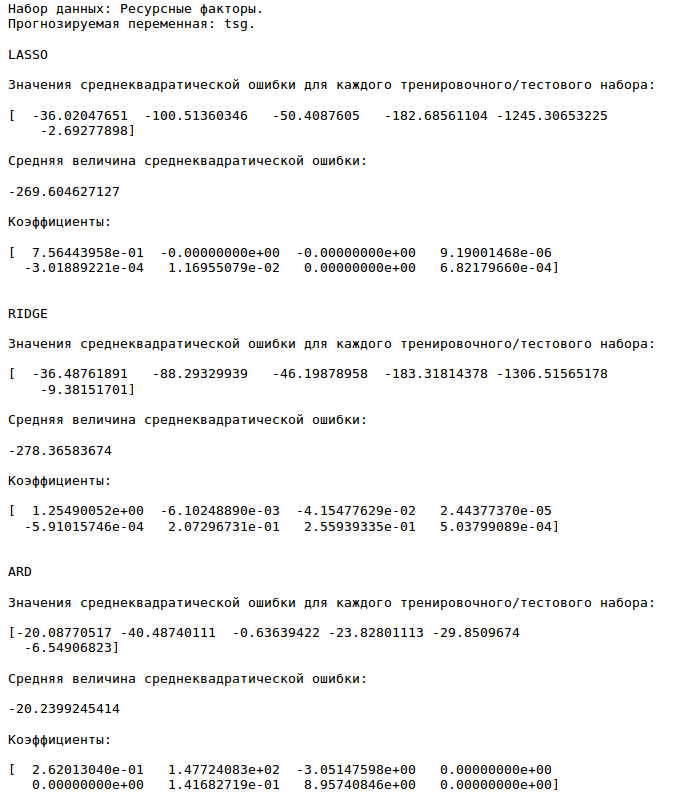


Рисунок А.5 — построение модели для числа зарегистрированных организаций ТОС по институциональным факторам.

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

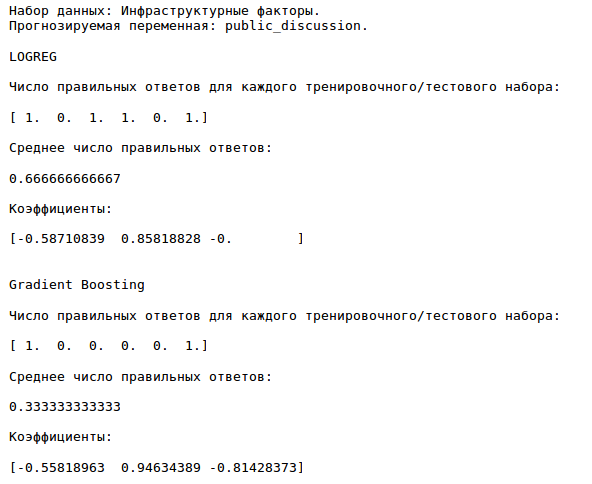
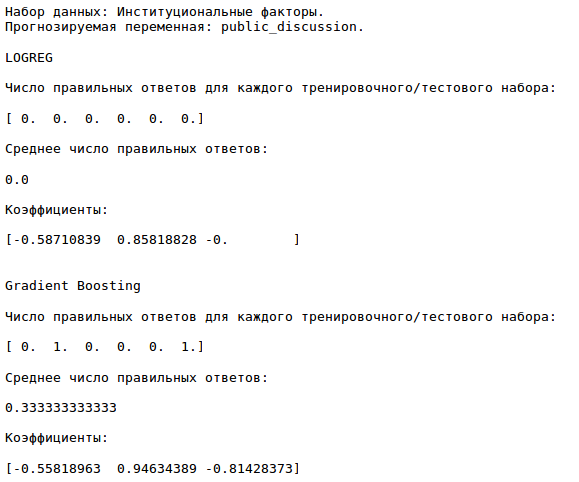
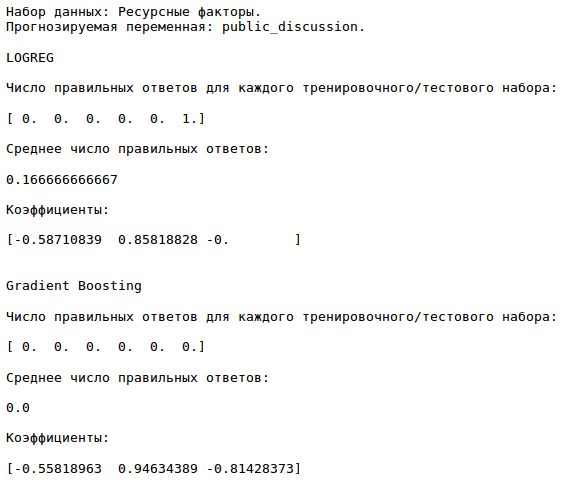


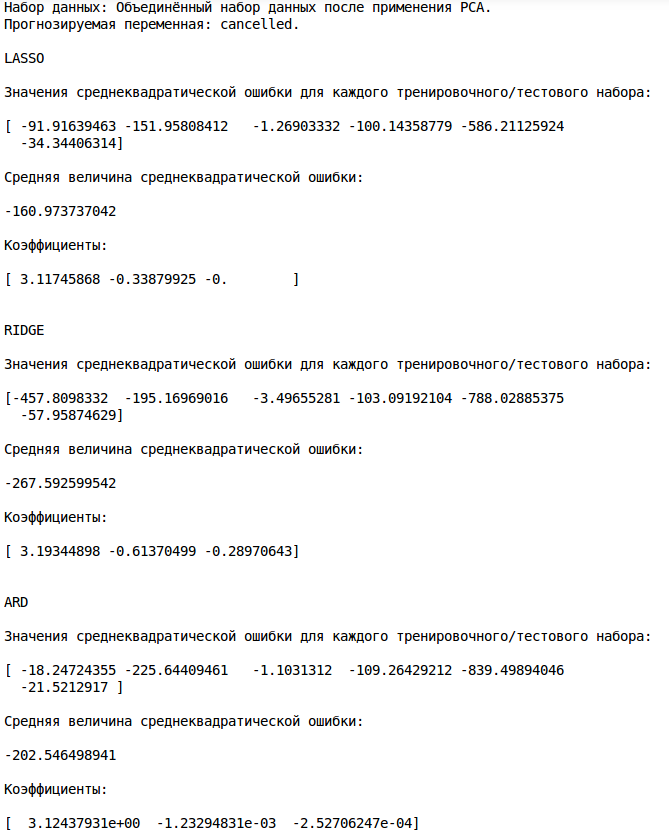
Рисунок Б.1 — построение модели для общественного обсуждения законопроектов в сети Интернет по инфраструктурным факторам.

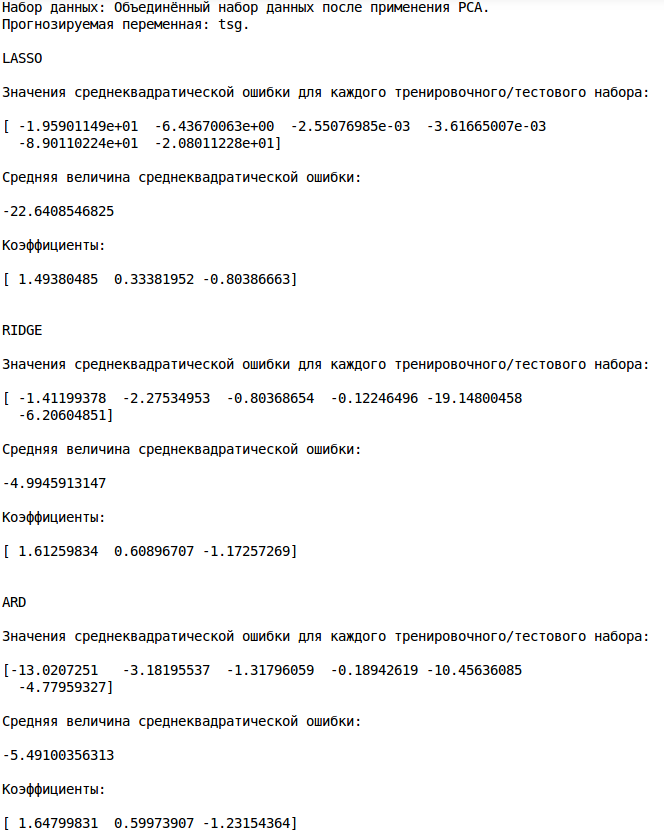
Рисунок Б.2 — построение модели для общественного обсуждения законопроектов в сети Интернет по институциональным факторам.

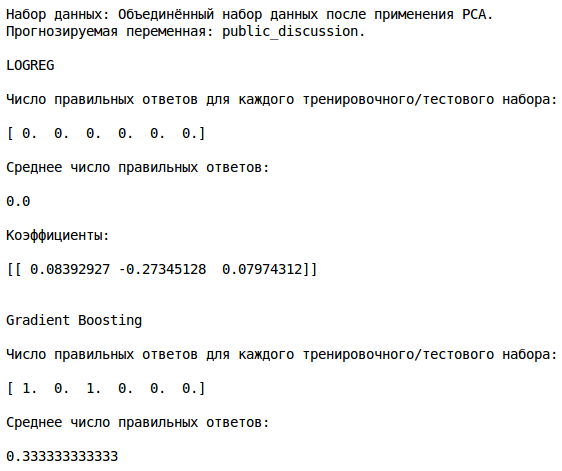
Рисунок Б.3 — построение модели для общественного обсуждения законопроектов в сети Интернет по ресурсным факторам.

ПРИЛОЖЕНИЕ В

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

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

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

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.