Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра Информатики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ _____
подпись инициалы, фамилия

« _____» _____ 2017 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 «Системный анализ и управление»

Непараметрический алгоритм прогнозирования стоимости автомобиля

Руководитель _____ ст. преподаватель, к.т.н. А. А. Корнеева
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ И.О.Бессмертный
подпись, дата инициалы, фамилия

Красноярск 2017

РЕФЕРАТ

Бакалаврская работа по теме «Непараметрический алгоритм прогнозирования стоимости автомобиля» содержит страницы текстового документа, 11 таблиц, 17 рисунков, 19 формул, 45 использованных источников.

КЛЮЧЕВЫЕ СЛОВА: ПРОГНОЗ, МАШИННОЕ ОБУЧЕНИЕ, ЛИНЕЙНАЯ РЕГРЕССИЯ, НЕПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ, КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, ИМПУТАЦИЯ, ОЦЕНКА НАДАРАЯ-ВАТСОНА, СОКРАЩЕНИЕ РАЗМЕРНОСТИ, AZURE.

Цель работы состоит в повышении точности решения задачи идентификации по многомерной выборке наблюдений с пропусками в условиях малого объёма данных.

Для достижении данной цели были поставлены следующие задачи:

1) исследование методов заполнения пропусков в данных и реализация наиболее подходящего из них;

2) реализация и исследование метода сокращения размерности;

3) синтез и исследование непараметрического алгоритма прогнозирования по многомерной выборке наблюдений.

Для решения поставленных задач в работе использовались методы статистического анализа, непараметрического оценивания, математического моделирования, анализа данных, сокращения размерности, заполнения пропусков.

В работе используются данные, представленные на репозитории https://archive.ics.uci.edu/ml/datasets/Automobile. Этот набор данных был создан американским статистом Джеффри Шлиммером на основе информации «Вашингтонского страхового института безопасности дорожного движения» для расчёта страхового риска автомобилей. Выборка состоит из 205 наблюдений, каждое из которых состоит из 26 признаков.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Информация играет важную роль в жизни общества. Каждому известно высказывание: "Кто владеет информацией, тот владеет миром!". И с ним невозможно не согласиться. Ещё с древних времен люди собирали и систематизировали информацию об окружающем мире, основывались в своих решениях на знаниях и опыте прошлых поколений, дополняли, обновляли их своими. Данные качества позволили человечеству выживать и развиваться. Чем больше проходило времени, тем большую роль начинала играть информация в жизни общества. Каждый раз достаточное её накопление давало толчок к развитию научно-технического прогресса.

В современном мире большие объёмы данных хранятся и анализируются в различных сферах жизни общества: в металлургии и менеджменте, ракетостроении и политологии, ботанике и банковском деле, экономике и медицине. Этот список можно продолжать бесконечно. Сегодня невозможно представить область науки или промышленную сферу, которая бы не имела собственную базу данных, собранную и "доведённую до совершенства" годами или десятилетиями кропотливого труда тысяч и миллионов человек.

Используя данную информацию, исследователи всегда хотели знать, как поведёт себя интересуемый объект при определённых условиях окружающего мира. С решением данного вопроса может помочь прогнозирование. Если говорить научным языком, то прогнозирование - это специальное научное исследование, обусловленное желанием знать поведение исследуемого процесса, построенное на основе имеющихся данных.

Но данные не всегда хранятся в виде, удобном для работы исследователя. Часто для решения поставленной задачи исследователь сталкивается с огромными объёмами лишних данных, которые будут лишь мешать достижению поставленной цели. Например, в медицине, если стоит задача определить, имеется ли у пациента рак, то данные о наличие у него плоскостопия или дефектов зрения не несут в себе никакой информативности, а будут лишь мешать

определению заболевания. Недостаток данных по объекту исследования так же, как и их излишек, является проблемой. Представим, что на заводе железобетонных конструкций имеется выборка наблюдений, содержащая в себе результаты измерений лишь 10 характеристик бетона марки B12 при разном соотношении воды, песка, гравия и цемента. Аналитику необходимо сделать прогноз качества бетона B12 при увеличении первых двух параметров и уменьшении оставшихся двух. Никакого приемлемого ответа на основе всего лишь 10 наблюдений он дать не сможет. Кроме того, в данных могут присутствовать выбросы и пропуски, с которыми исследователю также необходимо как-то бороться. Пропуски могут быть вызваны недобросовестным отношением к своей работе человека, отвечающего за сбор информации или же невозможностью измерения какого-либо параметра в конкретных условиях.

С аналогичной проблемой человек сталкивается при оценке стоимости автомобиля. В современной жизни автомобиль является неотъемлемой частью жизни каждого из нас. Наличие машины говорит об успешности человека. Личное транспортное средство помогает сохранить время на преодолении расстояний в условиях мегаполисов, почувствовать себя свободным и независимым. Многие люди считают, что автомобиль - это не роскошь, а средство передвижения. Но это не совсем так.

Перед приобретением нового автомобиля люди задают себе стандартные вопросы: "Стоит ли данная покупка тех денег, которые запрашивает за него автозавод?", "Не будет ли обслуживание купленного транспортного средства наносить серьёзный ущерб финансовому состоянию покупателя?". Ответ на эти вопросы может дать решение задачи прогнозирования.

1 Задача прогнозирования в условиях многомерной выборки, содержащей пропуски

1.1 Задача моделирования и прогнозирования

1.1.1 Моделирование

Моделирование - это научно обоснованный метод оценки характеристик сложных систем, необходимый для принятия решений в различных сферах жизни общества. Моделирование помогает эффективно исследовать реальные системы с помощью математических моделей, написанных на современных вычислительных машинах, которые являются инструментом в руках исследующего.

Как в основе любых открытий лежит воображение, так и в основе любого моделирования располагается теория подобия. Основная идея данной теории состоит в том, что абсолютное подобие имеет место лишь, когда происходит замена одного объекта точно таким же. При моделировании построить идеальную модель невозможно, поэтому стремятся достичь, чтобы модель с достаточной точностью отображала исследуемую сторону объекта.

Существует великое множество классификаций видов моделирования, ниже будут приведены только самые известные.

Степень полноты модели является одним из важнейших признаков классификации моделирования. Она показывает насколько хорошо модель описывает все стороны исследуемого объекта. Согласно данному признаку в работе [1] модели подразделяются на:

- полные;
- неполные;
- приближённые.

В зависимости от характера исследуемых объектов, виды моделирования по Самарскому А.А. [2] могут разделятся на:

- детерминированные или стохастические;
- статические или динамические;

- дискретные или непрерывные, или дискретно-непрерывные.

Детерминированное моделирование используется для отображения процессов, в которых отсутствуют любые случайные воздействия. Стохастическое моделирование позволяет работать с вероятностными процессами, оценивая средние характеристики.

Статическое моделирование описывает поведение объекта в определённый момент времени, а динамическое служит для описания развития и изменения системы во времени.

Дискретные модели применяются к системам, поведение которых изменяется лишь в заданные моменты времени, непрерывные модели используются для систем, поведение которых изменяется непрерывно во времени, дискретно-непрерывное используется в тех случаях, когда в исследуемом объекте присутствуют и дискретные, и непрерывные процессы.

В зависимости от формы представления модели Мухин О.И. [3] выделяет материальное и абстрактное моделирование. Абстрактное моделирование применяют в том случае, когда модель реального объекта невозможно по каким-либо причинам создать в данный момент времени. Материальное моделирование подразумевает создание материального аналога исследуемого объекта, воспроизводящего его основные характеристики.

Также в своей работе [4] Орлов А.И. делит модели по отрасли деятельности исследующего их человека, здесь выделяют модели: физические, математические, биологические, экономические политические и т.д.

Олзоева С.И. [5] считает, что при построении модели необходимо руководствоваться методом моделирования, который состоит из следующих этапов:

а) постановка целей и задач построения модели;

б) анализ создаваемой модели и определение области её применения;

в) практическое применение полученных данных;

г) корректировка созданной модели.

Одним из важных этапов разработки модели служит этап постановки задачи. На этом этапе необходимо поставить цель разработки модели, описать задачи и провести анализ моделируемого объекта. Задача - это проблема, для решения которой и строится модель. Чётко формулируется, что должно быть получено при решении поставленных задач. На этом этапе выделяются основные сущности, связи и процессы, протекающие в исследуемой системе и определяются границы рассмотрения модели.

Моделирование играет важную роль в такой отрасли современной науки, как машинное обучение. На основе методов машинного обучения строятся достаточно точные модели систем, которые исследователь не может построить, используя классическое моделирование, либо из-за неполного понимания предметной области, либо из-за протекания сложных процессов в этой системе.

1.1.2 Машинное обучение

Машинное обучение - это класс методов искусственного интеллекта, с помощью которых происходит не прямое решение конкретной задачи, а обучение решению на примере сходных задач. Данные методы основаны на средствах теории вероятностей, теории графов, математической статистики, теории алгоритмов, методов оптимизации и множестве численных методов.

Машинное обучение, как отдельное направление, сформировалось в конце 80-хх годов прошлого века в результате отделения от науки о нейросетях. Поэтому неудивительно, что ряд задач, решаемых машинным обучением, также решают и нейронные сети.

Цель машинного обучения состоит в частичной или полной автоматизации решений профессиональных задач в разнообразных областях жизни современного общества.

Сфера применения машинного обучения ежедневно расширяется. Происходит накопление огромных объёмов информации, хранящейся в базах знаний учреждений. Остро встаёт вопрос решения задач прогнозирования,

управления, обработки информации и принятия решений. Для нахождения оптимального решения обращаются к машинному обучению. Раньше, когда данных такого объёма не было, эти задачи либо не возникали, либо решались совершенно другими методиками.

К классическим задачам машинного обучения относят:

а) Обучение с учителем - задаётся пара "ситуация - решение":

а.1) Задача классификации. Имеется множество объектов X и множество номеров классов Y. Необходимо построить алгоритм, способный отнести произвольный объект $x \in X$ в некоторый класс $y \in Y$;

а.2) Задача восстановления регрессии. Имеется множество объектов X и множество номеров классов Y. Причём |У| = |R|, где R - множество действительных чисел. Необходимо построить алгоритм, способный отнести произвольный объект $x \in X$ к некоторому классу $y \in Y$;

а.3) Задача предсказания. Имеется множество X, которое является временным рядом. Необходимо найти значения функции за пределами X.

б) Обучение без учителя - задаётся только "ситуация", необходимо принять "решение":

б.1) Задача кластеризации. Имеется множество объектов X. Необходимо построить алгоритм, который сможет отнести произвольный объект $x \in X$ к некоторому кластеру(классу). Количество кластеров неизвестно;

б.2) Задача поиска ассоциативных правил. Имеется множество объектов X. Необходимо построить алгоритм, который сможет находить взаимосвязи между любыми элементами X;

б.3) Задача сокращения размерности данных. Имеется множество объектов X, являющееся декартовым произведением $Y \times Y \times ... \times Y$. Необходимо сократить размерность множества X с минимальной потерей качества данных;

б.4) Задача ранжирования. Имеется пара "запрос-объект". Необходимо определить соответствие объекта запросу.

Также стоит заметить, что методы машинного обучения не всегда решают какую-то одну из вышеперечисленных задач, намного чаще они находят решение сразу нескольких задач опираясь на методы математической статистики и других дисциплин.

Задача прогнозирования является самой популярной задачей машинного обучения. В данной работе, как главная задача машинного обучения, будет решаться задача а.2. Восстановление регрессии является наиболее подходящим методом для решения задачи прогнозирования.

1.1.3 Прогнозирование

Прогнозирование в переводе с греческого означает "знание наперёд". Если говорит научным языком, то прогнозирование - это специальное научное исследование, направленное на предсказание поведения процесса или системы в будущем.

Прогнозирование невозможно без такого термина, как "прогноз". Прогноз - это научно обоснованное суждение о состоянии системы в будущем, о возможных путях её развития и способах его достижения. Владимирова Л.П. [6] считает, что прогноз обязан удовлетворять некоторым принципам:

- в момент выдвижения прогноза нельзя однозначно сказать, является он истинным или ложным, так как прогноз высказывается ещё на непроизошедшее событие;

- он должен строго определять границы временного интервала, внутри которого должно будет произойти прогнозируемое событие;

- на момент высказывания прогноза прогнозирующий должен располагать информацией, с помощью которой он сможет оценить точность и надёжность прогноза.

Процесс разработки прогноза состоит в обработке информации об имеющемся объекте прогнозирования, для получения направления на его дальнейшее развитие. Главной задачей при этом является получение

объективной научно обоснованной оценки поведения прогнозируемой системы в будущем, опираясь на имеющуюся информацию об этой системе. А второстепенной, но от этого не менее важной, задачей прогнозирования можно назвать выявление значимых факторов, которые будут влиять на исследуемую систему в дальнейшем.

В своей работе [7] Л.Н. Слуцких выделяет следующие принципы, на которых базируется процесс прогнозирования:

- научная обоснованность прогноза - использование научных методов в его разработке, с учётом закономерностей развития прогнозированного объекта;

- непрерывность прогнозирования - при изменении поведения прогнозированного объекта, прогноз должен корректироваться;

- согласованность прогнозов - разработанный прогноз не должен кардинально отличаться от других прогнозов об этом объекте, полученных на основе тех же данных;

- многовариантность, альтернативность прогноза - разработка нескольких прогнозов для использования их в случае изменении поведения прогнозируемого объекта;

- выбор основных факторов - при прогнозировании должны учитываться все основные факторы, влияющие на исследуемый объект;

- системность разработки прогноза - с одной стороны, процесс прогнозирования следует рассматривать, как единую неделимую систему, а с другой стороны, как сложную систему, состоящую из отдельных частей;

- верифицированность прогноза - спрогнозированные оценки должны быть достоверными и обоснованными;

- адекватность - максимальное приближение спрогнозированной оценки к реальной действительности;

- рентабельность - эффект от прогноза должен превышать затраты на его разработку.

Принципы прогнозирования обеспечивают единство разнообразных методов и моделей прогнозирования. Они находят своё отражение в различных

тенденциях современного прогнозирования: интеграция прогнозирования в системы управления, увеличение локальных прогнозов по отношению к глобальным, повышение доли экспертных методов по сравнению с математическими и т.д.

В современном мире существует огромное разнообразие классификации прогнозов. Ниже представлена классификация, использованная в работе Поздеевой О.Г. и Новиковой О.В. [8].

По цели разработки выделяют нормативные и поисковые прогнозы. Поисковые выявляют будущее развитие исследуемой системы на основании тенденции прошлого, а нормативные разрабатываются с учётом заранее поставленных целей и сроков их достижения.

По временной тенденции существуют следующие виды прогнозов, представленные в таблице 1.

Таблица 1 - Классификация прогнозов по времени их действия

Название	**Срок разработки**	**Содержащиеся показатели**
Оперативные	до одного месяца	только количественные
Краткосрочные	до одного года	общие количественные
Среднесрочные	1-5 лет	количественные и общие качественные
Долгосрочные	5-15 лет	общие количественные и общие качественные
Дальнесрочные	свыше 15 лет	общие качественные

По методам разработки выделяют интуитивные и формализованные прогнозы. Интуитивные прогнозы строятся по информации, полученной от экспертов. Они используются, например, когда объект прогнозирования новый, и о нём нет информации, или когда объект прогнозирования сложный(на него оказывают влияние множества факторов). Формализованные основываются на фактической информации о прогнозируемом объекте.

1.2 Известные методы и алгоритмы

Идентификация – это определение структуры системы и её параметров на основе анализа входных и выходных переменных этой системы. Главной задачей идентификации является построение некоторой оптимальной модели на основе наблюдений над входными и выходными переменными [9]. Для формулировки задачи идентификации с математической точки зрения необходима априорная информация об объекте исследования, которая состоит из информации о случайных помехах, ограничениях и критерии оптимальности. Критерий оптимальности представляет собой требования, удовлетворение которых поможет достичь наилучшего результата, а ограничения показывают возможности модели.

Априорная информация основывается на различных физических, химических, механических, биологических и других процессах или результатах наблюдений, которые предшествовали изучению интересующего нас объекта [10]. Априорные сведения могут утратить свою достоверность с прошествием некоторого промежутка времени.

Следует не путать априорную и апостериорную информацию. Апостериорная или текущая информация накапливается в результате выполнения эксперимента и является выборкой "входных-выходных" переменных наблюдения. Апостериорная информация обновляется в каждый момент времени выполнения эксперимента. Иногда она используется для накопления априорной информации для данного объекта, но в основном текущая информация предназначена для компенсации недостатка априорной информации. Недаром Я.З. Цыпкин как-то сказал [11]: «Априорная информация – это основа для формулировки проблемы оптимальности. Текущая информация – средство решения этой проблемы».

Наличие полной априорной информации об объекте невозможно, так как в любом процессе присутствует множество случайных факторов. Таким образом, исследователь всегда работает только с неполной априорной информацией.

Выделяют следующие уровни априорной информации:

- байесов уровень. Известны: параметрическая модель объекта исследования, законы распределения помех и уравнения каналов связи. Необходимо оценить параметры параметрической модели объекта;

- уровень параметрической неопределённости. Известны параметры параметрической модели исследуемого объекта, которые необходимо оценить. Также известны некоторые характеристики случайных помех;

- уровень непараметрической неопределённости. Неизвестна параметрическая структура исследуемого объекта. Для решения задачи идентификации используют методы непараметрической статистики [12];

- уровень параметрической и непараметрической неопределённости. Имеется параметрическая и непараметрическая априорная информация. Необходимо оценить параметры модели, состоящей из параметрических и непараметрических соотношений.

Исходя из уровня априорной информации решают задачи идентификации в "узких" и "широких" смыслах.

Для моделирования различных дискретно-непрерывных процессов используют теорию идентификации в "узком" смысле [13]. Она говорит о том, что сначала, на основе имеющейся априорной информации, выделяется параметрический класс оператора A^{α}, примерный вид представлен в формуле (1):

$$x_{\alpha}(t) = A^{\alpha}(u(t), \alpha), \quad (1)$$

где $x_{\alpha}(t)$ - параметрическая модель;

A^{α} - параметрическая структура модели;

α - вектор параметров.

Далее происходит оценивание параметров α на основе используемой выборки. Чаще всего оценивание осуществляется методом стохастических аппроксимаций или методом наименьших квадратов. Решение задачи идентификации зависит от того, насколько хорошо выбран оператор (1).

Для идентификации в "широком" смысле отсутствует выбор параметрического класса оператора [11]. В работе [9] говорится, что в данном случае априорная информация об объекте либо полностью отсутствует, либо её недостаточно. По этой причине исследователю необходимо решить ряд дополнительных задач. Например, выбор структуры модели, оценивание стационарность и линейность объекта исследования, оценивание влияние входных переменных на выходные, отбор значащих переменных и др. Одним из методов решения задач идентификации в "широком" смысле является оценка функции регрессии [14].

1.2.1 Параметрические модели

Главными задачами параметрической идентификации считают определение структуры и параметров настраиваемой модели по наблюдаемым входным воздействиям и выходным величинам, которые обеспечивают экстремум выбранного критерия, описывающего качество идентификации [15]. Также считается, что известны параметры структуры объекта исследования.

Настройка параметров выбранной параметрической структуры модели описана в огромном количестве работ [11,9,16 и др.]. А вот трудов о выборе самой структуры почти нет. И это при том, что при параметрическом подходе именно выбор математического описания объекта играет важную роль в получении результата решения задачи идентификации.

Общая схема параметрической идентификации, представленная в работе [11], показана на рисунке 1. Здесь u_t – значения входных переменных процесса; x_t – значения выходных переменных; ξ_t – случайная помеха; x_{st} – выход настраиваемой модели; α – вектор параметров настраиваемой модели; I_s – вектор всех наблюдений; ε_t – ошибка рассогласования; $Q(\varepsilon)$ – функция потерь; M – математическое ожидание; $R(\alpha)$ – критерий идентификации.

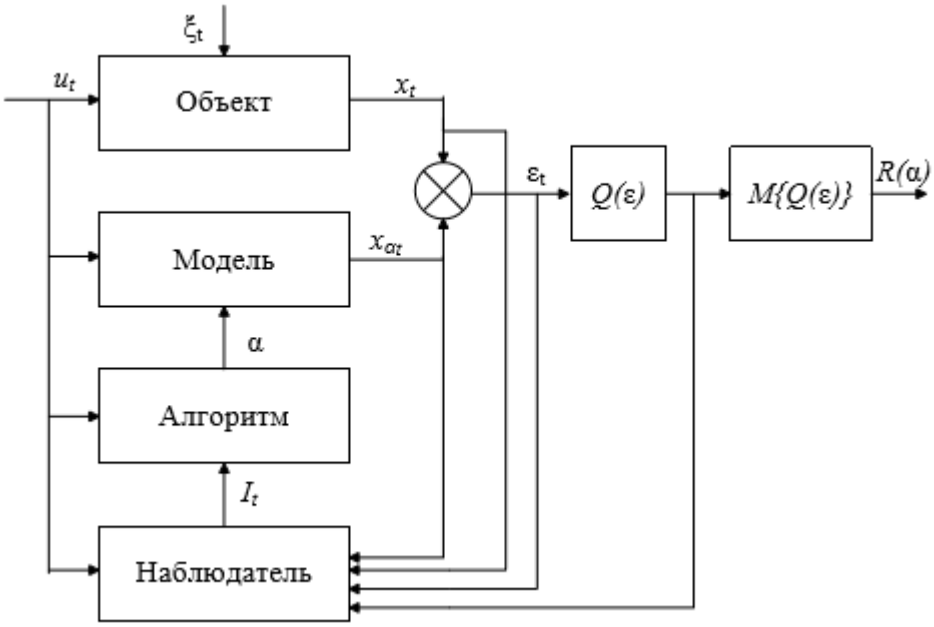

Рисунок 1 - Схема задачи идентификации

Данная схема говорит нам о том, что идентификация осуществляется с на основе настраиваемой модели, заданной параметрической структуры (блок "Модель"). Параметры модели α рассчитываются по поступающим наблюдениям I_s от блока "Наблюдатель" в блок "Алгоритм".

Соответствие модели объекту определяет критерий качества идентификации, представленный в формуле(2):

$$R(\alpha) = M\{Q(\varepsilon(x_t, x_{st}))\}. \quad (2)$$

Блок "Алгоритм" представляет из себя некоторый алгоритм идентификации, выбранный исследователем и позволяющий произвести оценку параметров модели α. Главная задача данного алгоритма состоит в минимизации критерия (2):

$$R(\alpha^*) = \min_{\alpha} R(\alpha). \quad (3)$$

Для настраивания параметров модели применяют разнообразные итеративные методы. Если функционал (3) дифференцируем, то он достигает своего экстремума при таких значениях $\alpha = (\alpha_1,..., \alpha_k)$, для которых k частных производных одновременно равны 0, что представлено в формуле (4):

$$\frac{\partial R}{\partial \alpha_j} = 0, j = \overline{1,k}. \quad (4)$$

Основная идея решения (3), описана в работе [13]. Преобразуем уравнение (4) к виду (5):

$$\alpha = \alpha - \gamma \nabla R(\alpha), \quad (5)$$

где γ - некоторой множитель, позволяющий найти оптимальный вектор $\alpha = \alpha^*$ с помощью последовательный приближений (6):

$$\alpha[n] = \alpha[n-1] - \gamma[n] \nabla_{\alpha} \mathrm{R}(\alpha[\mathrm{n}-1]). \quad (6)$$

Методы, основанные на использовании формулы (6) для нахождения α^*, известны в теории идентификации, как регулярные итеративные методы. Также для нахождения неизвестных параметров применяются методы стохаппроксимаций и метод наименьших квадратов.

1.2.2 Непараметрические модели

Для работы методам параметрической идентификации требуется большое количество априорной информации, необходимой для определения структуры исследуемого объекта. Но чаще всего бывает, что априорной информации об объекте недостаточно и невозможно определить структуру объекта с требуемой

точностью. П. Эйкхофф в своей работе [9] «Значение структуры нельзя переоценить. Ее выбор определяется типом применения модели и может оказаться решающим фактором успеха или неудачи принятой схемы оценивания». В условиях недостаточности априорной информации разумно использовать методы непараметрической идентификации [17,18].

В работах [19,20,21] для построения моделей в условиях непараметрической неопределённости рассчитывается непараметрическая оценка кривой регрессии, вид которой представлен в формуле (7):

$$x_s(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}, \quad (7)$$

где $u = (u_1, u_2, ..., u_m)$ - m-мерный вектор входных воздействий;

x - выходная переменная;

$\Phi(c_s^{-1}(u-u_i))$ - ядерная колокообразная функция;

c_s - коэффициент размытости ядра.

Ядерная функция и коэффициент размытости ядра должны удовлетворять условиям сходимости, представленным в работе [22] и показанным в формуле (8):

$$\begin{aligned} &c_s > 0; && \Phi\left(c_s^{-1}(u-u_i)\right) < \infty; \\ &\lim_{s\to\infty} c_s = 0; && c_s^{-1} \int_{\cap(u)}^{\infty} \Phi\left(c_s^{-1}(u-u_i)\right)\, dx = 1; \\ &\lim_{s\to\infty} s c_s^m = \infty; && \lim_{s\to\infty} c_s^{-1}\ \Phi\left(c_s^{-1}(u-u_i)\right) = \delta(u-u_i), \end{aligned} \quad (8)$$

где $\delta(u-u_i)$ - дельта-функция Дирака.

В качестве ядерных функций используются функции, содержащие различные ядра. Далее будут приведены некоторые из них.

Треугольное ядро представлено в формуле (9):

$$\Phi\left(\frac{x-x_i}{c_s}\right) = \begin{cases} 1 - |c_s^{-1}(x - x_i)|, & |c_s^{-1}(x - x_i)| \leq 1; \\ 0, & |c_s^{-1}(x - x_i)| > 1; \end{cases} \quad (9)$$

Параболическое ядро представлено в формуле (10):

$$\Phi\left(\frac{x-x_i}{c_s}\right) = \begin{cases} 0.75(1 - (c_s^{-1}(x - x_i))^2), |c_s^{-1}(x - x_i)| \leq 1; \\ 0, & |c_s^{-1}(x - x_i)| > 1; \end{cases} \quad (10)$$

Кубическое ядро представлено в формуле (11):

$$\Phi\left(\frac{x-x_i}{c_s}\right) = \begin{cases} (1 + 2|c_s^{-1}(x - x_i)|)(1 - (c_s^{-1}(x - x_i))^2), |c_s^{-1}(x - x_i)| \leq 1; \\ 0, & |c_s^{-1}(x - x_i)| > 1. \end{cases} \quad (11)$$

На рисунке 2 показана графическая интерпретация вышеописанных ядер:

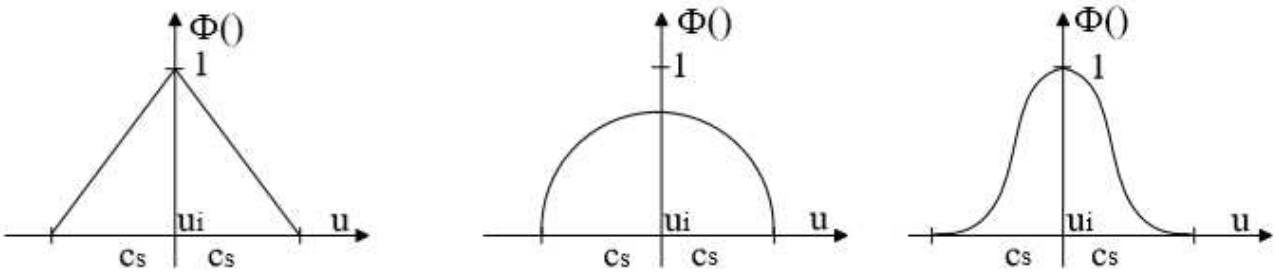

Рисунок 2 - Виды ядерных функций

Далее, для простоты, будем обозначать ядерную функцию как Φ(•). Обычно выбор формы ядра не сильно влияет на точность восстановления функции регрессии, в выборе Φ(•) исследователь основывается на собственный опыт. Выбор формы ядра может зависеть от дополнительных условий, например, требованиями дифференцирования.

Куда большую значимость на функцию качества оказывает коэффициент размытости ядра c_s - это константа, от величины которой зависит степень

"размытости" дельта-функции в окрестностях каждой точки выборки и степень гладкости рассчитанной оценки.

Коэффициент размытости c_s определяется решением задачи минимизации квадратичного показателя, показывающего соответствие выхода модели к выходу объекта, опирающегося на метод скользящего экзамена, приведённого в формуле (12):

$$R(c_s) = \sum_{k=1}^{s} (x_k - x_s(u_k, c_s))^2 = \min, k \neq i. \tag{12}$$

Непараметрическую оценку функции регрессии (7) можно привести к виду, показанному в формуле (13):

$$x_s(u) = \sum_{i=1}^{s} x_i \varphi(u, u_i), \tag{13}$$

где

$$\varphi(u, u_i) = \frac{\prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}. \tag{14}$$

1.3 Заполнение пропусков в матрице наблюдений

На сегодняшний день человечество оперирует с огромными объёмами информации. Но эта информация не всегда является полной. В ней возможны пропуски данных, допущенные, возможно, по вине собирающего эту информацию или отсутствии необходимых данных. Наличие пропусков, как и работа только с полными наблюдениями (после удаления наблюдений,

содержащих пропуски) может привести к неточным результатам и к искажению выводов, которые будут приняты в результате исследования.

Данная проблема может быть решена различными способами. Одни просто исключают наблюдения с пропусками. Другие стремятся на этапе первичной обработки восстановить исходную зависимость одним из методов заполнения пропусков в матрице наблюдений. Ниже будет рассмотрен именно второй вариант.

1.3.1 Известные методы

Известно много способов заполнения пропусков уже после этапа сбора информации: заполнение средним значением, заполнение модой, заполнение значениями предыдущих наблюдений, заполнение на основе градации шкалы, расчёт возможных значений при построении регрессионной модели, восстановление пропущенного значения сплайн-интерполяцией и многие другие. Заполнение пропусков необходимо не только для получения дополнительной информации, но и для сохранения уже имеющейся, за счёт оставления наблюдений, содержащих пропуски.

Но заполнение пропусков, или импутирование, имеет и ряд недостатков, которые невозможно игнорировать [23]:

а) искажение структуры результирующих данных;

б) смещение полученного результата от реального.

Вероятно, что такие модели будут менее точными, если сравнивать с моделями, построенными только на полных наблюдениях. Потери в точности зависят от выбранного метода импутирования и от качества предсказания недостающих значений. Выбор конкретного метода заполнения пропусков должен зависеть от метода анализа данных, который будет применятся в дальнейшем.

Знание механизма, который приводит к отсутствию значений, сильно влияет на выбор метода анализа и интерпретации результатов. Многие методы

обработки пропусков данный механизм образования пропусков явно не включают. Однако им нельзя просто пренебречь, так как, например, отказ от ответа в опросе о доходах скорее всего подразумевает наличие тайных доходов, что нельзя игнорировать.

Наиболее распространённая классификация методов импутирования, созданная таким известным статистом, как Литтл Р.Д. [24], отражена на рисунке 3.

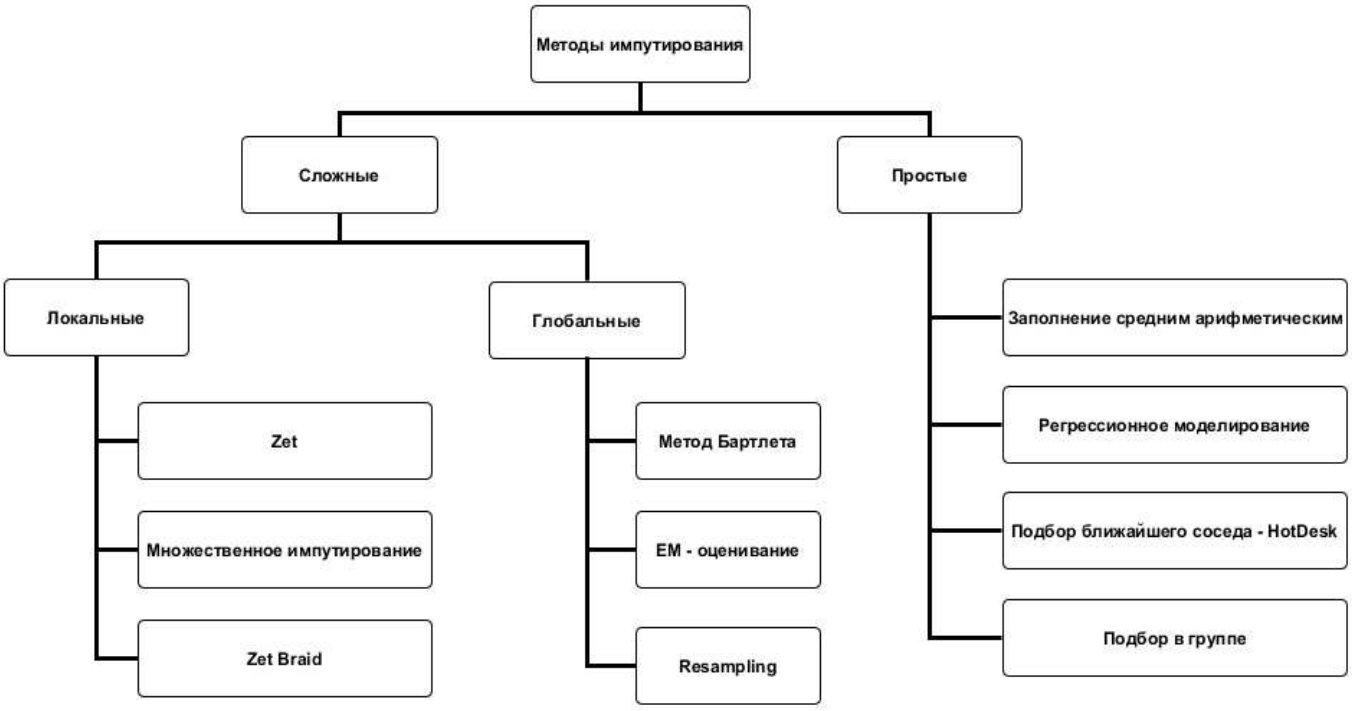

Рисунок 3 - Классификация методов импутирования

Простые методы - это неитеративные методы, основанные на простейших арифметических операциях.

Сложные методы - это итеративные методы, основанные на оценки точности подставляемого вместо пропуска значения. Они подразделяются на локальные и глобальные.

Глобальные методы - это методы, в которых в оценивании любого из пропущенных значений участвуют все наблюдения рассматриваемой выборки.

Локальные методы - это методы, в которых в оценивании любого из пропущенных значений участвуют наблюдения без пропусков, находящиеся в некоторой окрестности рассматриваемого наблюдения.

Рассмотрим каждый представленный метод в отдельности. Первые два метода хорошо представлены в работе [25].

Заполнения средним. Самый простой метод заполнения пропусков из существующих. Суть метода состоит в том, чтобы все пропущенные значения наблюдений заменить на среднее арифметическое, высчитанное на основе оставшихся, заполненных наблюдений.

Данный метод реализуется довольно просто, но обладает рядом нежелательных свойств:

а) происходит занижение оценки истиной дисперсии из-за заниженного объёма выборки.

б) невозможность оценивания корреляции между двумя переменными с помощью стандартных методов.

в) искажение распределения значений, что является важным недостатком для наблюдателя, исследующего выборку по графикам или гистограммам.

Подбор внутри групп. Данную процедуру можно описать как метод, в котором происходит выбор группы по определённому признаку для каждого пропущенного значения. При заполнении с подбором выбираются различные значения из наблюдений без пропусков сходных с данным. Данный метод широко распространен, имеет множество вариаций, использующих очень сложные схемы отбора групп. Недостатком метода является искажение распределения обрабатываемой выборки.

Метод Hot Desk. Данный метод представлен в работе [26] и представляет собой замену пропущенного значения на значение этой переменной у наиболее близкого объекта с полной информацией. Подбор ближайшего объекта может осуществляться как из всех полных наблюдений, так из небольшой подгруппы, выбранной по какому-либо признаку.

Для заполнения пропуска рассчитывается расстояние до всех полных объектов матрицы наблюдений и пропущенное значение по данной переменной заполняется таким же значением, как и у объекта, до которого расстояние является минимальным.

Тип функции расстояний рассчитывается в соответствии с типом использующихся данных, из характера связи между переменными и задач, поставленных исследователем.

Главным недостатком данного метода является появление зависимостей между восстановленными значениями и занижение оценки дисперсии. Снизить уровень зависимостей возможно только при большом числе подгрупп, что в свою очередь подразумевает большой объём выборки. Несмотря на этот недостаток метод hot desk применяется в работе большинства международных статистических организаций.

Метод Бартлета. Злоба Е. и Яцкив И. считают [25], что этот метод состоит из двух этапов:

a) подстановка вместо пропусков начальных значений;

b) проведение ковариационного анализа искомой переменной.

На втором этапе используется индикатор полноты наблюдений, он показывает, имеется ли в матрице наблюдений пропущенное значение. Индикатор полноты наблюдений равен 0, если значение не является пропуском, или равен 1, если значение пропущено.

Метод имеет следующие преимущества:

- он неитеративный;

- если структура пропусков является вырожденной, то этот метод "предупреждает" исследователя, что ответ будет некорректен.

Этот метод является привлекательным, но часто его нельзя реализовать непосредственно, так как многие программы дисперсионного анализа не имеют возможности работать при большом количестве сопутствующих переменных.

Метод Zet. Данный метод, а также его модернизация, описанная ниже, хорошо представлена в статье [27]. Суть метода Zet состоит в том, что при

подборе каждого пропущенного значения используются не все полные наблюдения, а только некоторая их часть, называемая компонентной матрицей. Данная матрица состоит из компонентных строк и столбцов. Компонентность объекта представляет собой некоторую величину, которая обратнопропорциональна декартовому расстоянию до строки - неполного наблюдения.

Далее строится функциональная зависимость прогнозируемого значения от значения взятого из компонентной матрицы. На основе зависимости происходит прогнозирование каждого пропущенного значения.

У метода Zet имеется 2 существенных недостатка:

а) в компонентную матрицу могут попасть неинформативные строки и столбцы, которые будут вносить помехи в прогнозируемые значения;

б) фиксированность размера компонентной матрицы.

Метод ZetBraid. Как не трудно догадаться метод ZetBraid является модернизированным методом Zet. Главной его отличительной чертой и достоинством является изменение размерности компонентной матрицы в ходе работы алгоритма.

В процессе работы производится отбор компонентных строк и столбцов. При каждом новом отборе создаётся новая компонентная матрица. По определённому критерию происходит определение её эффективности при прогнозировании пропущенного значения.

Метод ZedBraid имеет один существенный недостаток - это расчёт статистической оценки неизвестного значения возможен только на основе корреляционно-регрессивного анализа.

Метод Resampling. Данный метод описан в работе [25] и является итеративным. Его работа состоит в замене наблюдений, содержащих пропуски, случайными строками, подобранными из матрицы полных наблюдений. После по регрессионному уравнению происходит предсказание отсутствующего значения.

Построение регрессионной модели повторяется заданное количество раз. После этого значения регрессионных коэффициентов усредняется и получают значения для заполнения пропусков, которые имеют максимальную точность прогноза недостающих значений.

Информационная избыточность на фоне малой мощности множества комплектных данных и информационная недостаточность при случайном формировании искомой характеристики не позволяют получать приемлемые результаты. Также отсутствует возможность оптимизации метода.

Множественное импутирование. Метод множественного импутирования был предложен в 1970 году Дональдом Рубиным и дополнен в его работе [28]. Суть данного метода состоит в том, что каждому пропуску приписывается несколько возможных значений. В настоящее время этот метод приобретает всё большую популярность, но реализован в основном в коммерческом программном обеспечении.

На место каждого пропуска подставляется сразу несколько возможных значений, разброс между которыми существенен, это говорит о неопределённости самой модели.

Данные каждого набора заполненных значений хранятся в собственном массиве, каждый из которых в последствии анализируется, как матрица наблюдений, не имеющая пропусков.

Недостатком метода является избыточность второстепенной информации.

Метод EM-оценивание. Теоретический обзор данного метода представлен Королёвым В.Ю. [29]. Он является универсальным методом максимизации ожиданий (EM - expectation maximization). С его помощью можно не только восстанавливать пропуски, но и оценивать средние значения, корреляционные и ковариационные матрицы.

Алгоритм EM-оценивания предназначен для решения задач оптимизации некоторого функционала, через поиск экстремума целевой функции.

Этот алгоритм работает в 2 этапа, названные по первым буквам аббревиатуры метода.

1) Этап E.

На первом этапе для пропусков рассчитываются ожидаемые значения целевой переменной на основе полных наблюдений. Далее, когда будет получен массив без пропусков, происходит оценивание основных статических параметров: показателя разброса, коэффициентов взаимной корреляции и ковариации переменных.

2) Этап M.

На втором этапе происходит максимизация степени ожидаемых и полученных данных, а также соответствию структуры заполненных данных структуре данных полных наблюдений.

Применение данного алгоритма осложняется, если имеется большое число пропущенных значений признака, кроме того больших вычислительных ресурсов.

Регрессионное моделирование пропусков. В работе [30] представлено моделирование пропусков на основе регрессии. Обычно, импутирование на основе регрессионной модели производится в два этапа:

1) Строится регрессионная модель по полным наблюдениям, оцениваются коэффициенты в уравнении, где зависимой переменной является переменная, в которой присутствуют пропущенные значения.

2) По полученному уравнению находится отсутствующее значение. В случае, если переменная является количественной, то на выходе алгоритма будет число, если качественной, то категория, к которой будет отнесён объект.

Регрессионная модель выбирается на основе уровня измерения зависимой переменной и независимых переменных, переменных, используемых для прогнозирования отсутствующих значений.

Метод невозможно применить, если в наблюдении имеется пропусков больше одного, так как это приводит к множеству решений. Ещё одной проблемой данного метода является невысокая точность заполнения при нелинейных зависимостях.

1.4 Сокращение размерности в многомерном пространстве данных

Методы снижения размерности часто применяются исследователями для решения конкретных прикладных задач. Ниже будут представлены методы, получившие широкую популярность в различных сферах научной работы.

В многомерном статистическом анализе любой объект представлен вектором произвольного размера. Воспринимать данные легче, когда они представлены числом или как точки на плоскости. Работать со скоплением точек в трёхмерном пространстве намного сложнее. А если размерность будет четырёхмерной? Пятимерной? N-мерной? Воспринимать такие данные становится невозможно. Поэтому, вполне логично, когда исследователь хочет перейти от многомерного пространства данных к выборке небольшой размерности.

Целью уменьшения размерности матрицы наблюдений является удаление факторов или признаков, от которых исследуемая переменная не зависит. Они ухудшают свойства статистических процедур, повышают дисперсию оценок параметров.

Задача сокращения размерности имеет следующую формулировку: «Имеется многомерная выборка. Требуется перейти от неё к совокупности векторов меньшей размерности, сохранив при этом структуру исходных данных, снизив до минимума потерю информации, содержащейся в данных.»

Ниже будут представлены наиболее перспективные методы снижения размерности.

1.4.1 Метод главных компонент

Данный метод является самым часто используемым методом сокращения размерности. Он состоит в последовательном выявлении направлений, где данные имеют наибольший разброс. В работе [31] дана постановка задачи: пусть

выборка представляет собой набор векторов, одинаково распределённых с вектором X = (x(1),x(2),...,x(n)). Тогда линейные комбинации показаны в формуле (15)

$$Y(\lambda(1), \lambda(2), \ldots, \lambda(n)) = \lambda(1)x(1) + \lambda(2)x(2) + \ldots + \lambda(n)x(n), \quad (15)$$

где

$$\lambda^2(1) + \lambda^2(2) + \ldots + \lambda^2(n) = 1 \quad (16)$$

Вектор (16) лежит на единичной сфере n-мерного пространства.

В труде Орлова А.И. [32] говорится, что метод главных компонент позволяет найти направление максимального разброса. Это есть ни что иное как λ, при котором дисперсия случайной величины Y(λ(1), λ(2), …, λ(n)) является максимальной. При этом вектор λ задаёт первую главную компоненту, а величина Y(λ) - это проекция случайного вектора X на ось первой главной компоненты. Далее рассматривается плоскость в n-мерном пространстве, которая перпендикулярна первой главной компоненте. На неё проектируются все элементы выборки. При этом происходит сокращение исходного пространства на единицу.

На следующей итерации процедура повторяется. В ней также находится направление наибольшего разброса, которое уже является второй главной компонентой. Определяется плоскость, перпендикулярная первым двум главным компонентам. На неё проецируются все точки исходной выборки. Размерность сокращается ещё на единицу и уже становится на 2 меньше, чем размерность начального пространства. Начинается следующая итерация.

Иными словами происходит построение нового базиса в n-мерном пространстве, ортами которого являются главные компоненты.

С каждой итерацией происходит уменьшение дисперсии. Алгоритм заканчивает работу, когда дисперсия становится меньше определённого

заданного порога. Если было отобрано при этом k главных компонент, то считается, что был выполнен переход от n-мерного пространства к k-мерному.

Метод главных компонент позволяет сократить исходное количество признаков и при этом не исказить структуру данных.

1.4.2 Факторный анализ

Предыдущий метод является одним из методов факторного анализа [33]. Множество алгоритмов факторного анализа объединены тем , что в каждом из этих алгоритмов происходит переход к новому базису в начальном n-мерном пространстве.

Главным отличием по сравнению с методом главных компонент является разбиение факторов на группы на основе нагрузок. В одну группу объединяют факторы, которые оказывают сходное влияние на элементы нового базиса. После того, как группы будут созданы, необходимо оставить по одному признаку из каждой группы. В некоторых модернизированных алгоритмах факторного анализа вместо отбора признака происходит формирование нового фактора, который станет центральным для рассматриваемой группы. Сокращение размерности происходит при отборе единичных представителей групп, при этом все остальные признаки считаются несущественными и исключаются из выборки.

Данная процедура осуществляется не только с помощью факторного анализа. Этим также занимаются алгоритмы кластер-анализа. Эти алгоритмы были рассмотрены в работах [34-36]. Основной идеей кластер-анализа служит введения меры близости между признаками. Мера близости - это величина, определённая на паре объектов, которая измеряет похожесть этих двух объектов.

1.4.3 Многомерное шкалирование

На основе мер близости между признаками создан огромный класс методов многомерного шкалирования. [37,38]. Они основаны на представлении любого наблюдения исходной выборки точкой геометрического пространства, координатами которой будут являться факторы, которые вместе достаточно адекватно описывают объект. Отношения между объектами заменяют на отношения между точками.

Данный метод призван не только выявить значимые характеристики, но и наглядно интерпретировать результаты. Он состоит из нескольких этапах:

а) создание матрицы попарных различий или матрицы субъективных предпочтений;

б) решение задачи построения координатного пространства и размещение на нём точек-объектов;

в) анализ и интерпретация полученных результатов.

Для определения искомого координатного пространства используются методы линейной и нелинейной оптимизации. Вводится понятие стресса - это критерий отображения, который измеряет расхождения между близостями. Определяется такое множество точек, которое даёт минимальное значение стресса. Значения координат этих точек и являются решением задачи. Формальным критерием адекватности решения может быть использован коэффициент корреляции между исходными и результирующими данными. Чем он выше тем адекватнее решение было найдено.

Выводы к главе 1

В результате написания первой главы можно сделать следующие выводы:

1) были получены необходимые теоретические знания в таких научных направлениях, как моделирование, прогнозирование и машинное обучение;

2) проанализированы возможные решения задачи идентификации на основе параметрического и непараметрического подходов;
3) рассмотрены существующие методы заполнения пропусков в матрице наблюдений, их достоинства и недостатки;
4) исследованы методы сокращения размерности в условиях наличия многомерной выборки.

На основе проведённых исследованиях будет получено решение задачи прогнозирования в многомерном пространстве.

2 Задача прогнозирования в многомерном пространстве

2.1 Задача прогнозирования в многомерном пространстве

2.1.1 Описание исходных данных

В работе используются данные, представленные на репозитории https://archive.ics.uci.edu/ml/datasets/Automobile. Этот набор данных состоит из 205 наблюдений, каждое из которых состоит из 26 признаков. Признаки можно разделить на 3 группы:

- спецификация автомобиля на основе различных характеристик. Данная группа содержит 10 категориальных переменных(марка автомобиля, тип топлива, трансмиссия, количество дверей, тип кузова, ведущие колёса, расположение двигателя, тип двигателя, количество цилиндров, топливная система) и 14 непрерывных переменных(колёсная база, длина, ширина, высота, масса, размер двигателя, диаметр выхлопной трубы, ход поршня, коэффициент компрессии, мощность двигателя, обороты двигателя, расход топлива в городских условиях, расход топлива на трассе, цена);

- нормированные потери автомобиля. Эта группа определяет количество денег, потраченных на обслуживание владельцем автомобиля в год. Это значение нормируется для всех автомобилей в рамках определённой классификации.

- рейтинг страхового риска. Данная величина показывает степень безопасного вождения автомобиля, она варьируется от -3 до 3, цифра 3, например, показывает, что автомобиль является довольно безопасным.

Признаки, а также принимаемые ими значения приведены в таблице 2.

Таблица 2 - Информация о признаках

№	Название признака	Диапазон признака
1	Рейтинг (x1)	-3, -2, -1, 0, 1, 2, 3.
2	Нормированные потери (x2)	от 65 до 256.
3	Марка автомобиля (x3)	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo.
4	Тип топлива (x4)	diesel, gas.
5	Трансмиссия (x5)	std, turbo.
6	Количество дверей (x6)	four, two.
7	Тип кузова (x7)	hardtop, wagon, sedan, hatchback, convertible.
8	Ведущие колёса (x8)	4wd, fwd, rwd.
9	Расположение двигателя(x9)	front, rear.
10	Колёсная база (x10)	от 86,6 до 120,9.
11	Длина (x11)	от 141,1 до 208,1.
12	Ширина (x12)	от 60,3 до 72,3.
13	Высота (x13)	от 47,8 до 59,8.
14	Масса (x14)	от 1488 до 4066.
15	Тип двигателя (x15)	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16	Количество цилиндров (x16)	eight, five, four, six, three, twelve, two.
17	Размер двигателя (x17)	от 61 до 326.
18	Топливная система (x18)	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19	Диаметр выхлопной трубы (x19)	от 2,54 до 3,94.
20	Ход поршня (x20)	от 2,07 до 4,17.
21	Коэффициент компрессии (x21)	от 7 до 23.
22	Мощность двигателя (x22)	от 48 до 288.
23	Обороты двигателя (x23)	от 4150 до 6600.
24	Расход топлива в городских (x24) условиях	от 13 до 49.
25	Расход топлива на трассе (x25)	от 16 до 54.
26	Цена (y)	от 5118 до 45400.

Набор данных содержит в себе пропуски. Количество пропусков для разных признаков приведена в таблице 3.

Таблица 3 - Пропуски в данных

Название признака	**Количество пропусков**
Нормированные потери	41
Количество дверей	2
Диаметр выхлопной трубы	4
Ход поршня	4
Мощность двигателя	2
Обороты двигателя	2
Цена	4
Итого	59

Кроме того 46 наблюдений из 205 имеют пропуск хотя бы в одном значении признака. Их них 11 наблюдений имеют по 2 пропуска и 2 наблюдения по 3 пропуска.

Рассмотрим в отдельности переменные каждого типа и проведем их краткий анализ.

2.1.1.1 Категориальные переменные

1) *Марка автомобиля*. Автомобили марок Toyota, Nissan и Mazda возглавляют рейтинг по производству автомобилей во всём мире. Неудивительно, что в репозитории такие автомобили составляют большинство. Данную тенденцию можно посмотреть на рисунке 4.

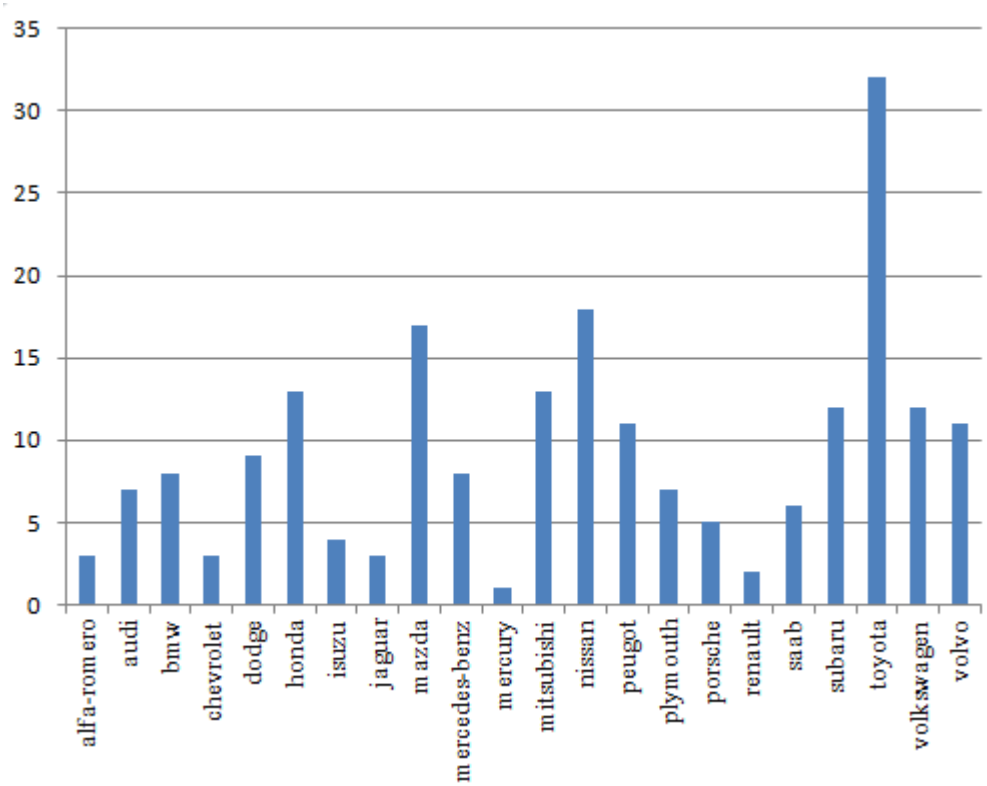

Рисунок 4 - Гистограмма марок автомобилей

2) *Тип топлива*. В наборе данных имеется всего 20 автомобилей, которые используют в качестве топлива дизель, остальным 185 транспортным средствам нужен бензин.

3) *Трансмиссия*. 37 автомобилей оснащены коробкой-автомат, 168 имеют механическую коробку передач.

4) *Количество дверей.* Имеется 89 двухдверных и 114 четырёхдверных автомобиля.

5) *Тип кузова.* Большинство автомашин - это sedan(96) и hatchback(70). Это хорошо видно на рисунке 5.

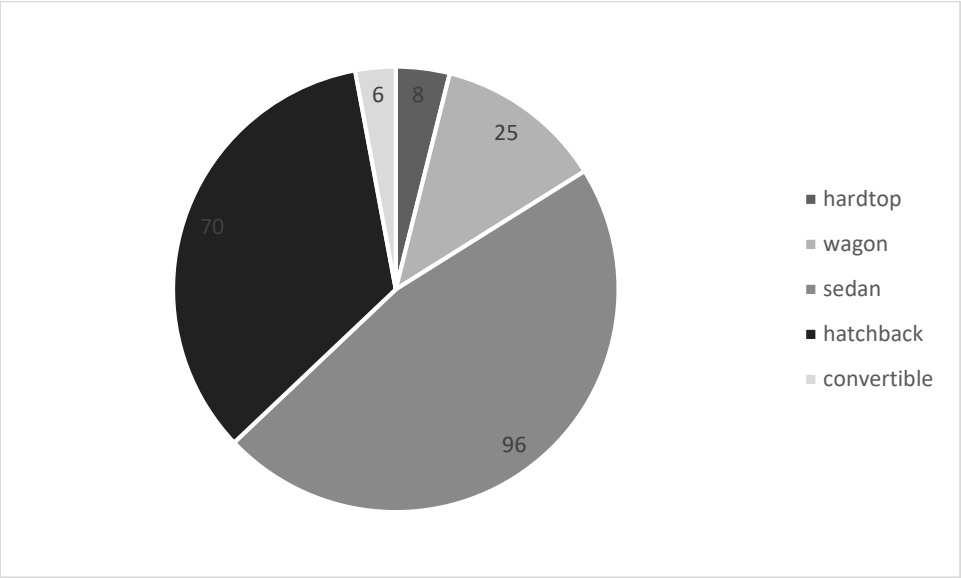

Рисунок 5 - Диаграмма автомобилей, с определённым типом кузова

6) *Ведущие колёса.* 120 транспортных средств имеют передний привод, 76 - задний привод и 9 - полный привод.

7) *Расположение двигателя.* На преобладающем большинстве автомобилей(202) двигатель расположен спереди, и только у 3 машин - сзади.

8) *Тип двигателя.* Преобладающая часть автомобилей снаряжены ohc-двигателем(148), что показано на рисунке 6.

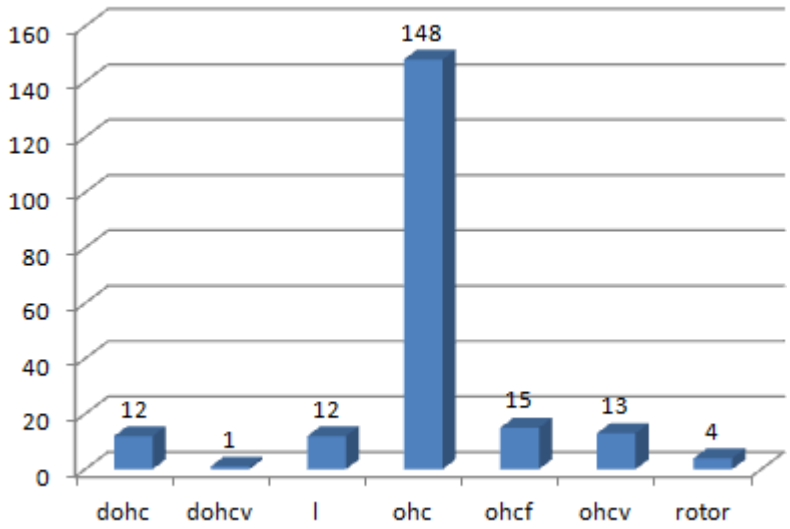

Рисунок 6 - Гистограмма типов двигателей

9) *Количество цилиндров.* В основном в выборке представлены автомобили с 4 цилиндрами(159). Рисунок 7 подтверждает данное наблюдение.

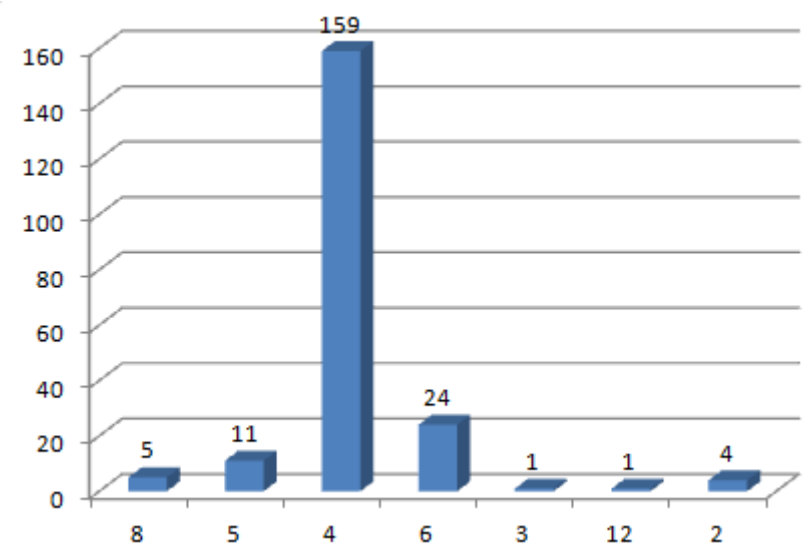

Рисунок 7 - Гистограмма количества цилиндров

10) *Топливная система.* Основными топливными системами являются 2bbl(66) и mpfi(94). Это показано на рисунке 8.

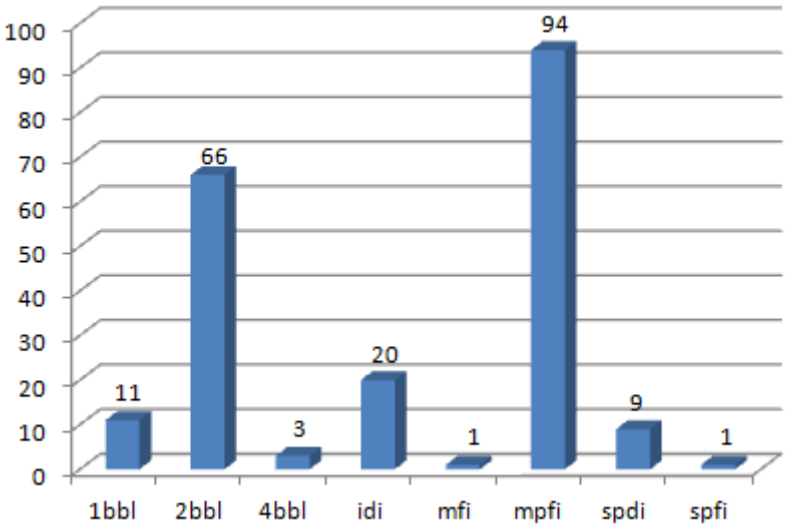

Рисунок 8 - Гистограмма топливной системы исходной выборки

2.1.1.2 Непрерывные переменные

Чтобы проанализировать непрерывные переменные, сначала необходимо выбрать метод, показывающий наличие зависимостей между данными переменными. Самым простым и эффективным из таких методов является корреляционный анализ. Коэффициент корреляции между двумя непрерывными переменными рассчитывался по формуле (17):

$$r_{x,y} = \frac{1}{s}\sum \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y}, \quad (17)$$

где, s - объём выборки;

x,y - непрерывные переменные;

m_x, m_y - математические ожидания непрерывных переменных x и y;

σ_x, σ_y - среднеквадратические отклонения непрерывных переменных x и y.

Проведя вышеупомянутый метод, нами было замечено, что цена сильно коррелирует с шириной, массой и объёмом двигателя автомобиля с корреляцией 0,84, 0,89 и 0,84 соответственно. Масса относится к весу автомобиля со всем стандартным оборудованием, но без пассажиров и груза. Размер двигателя влияет на мощность автомобиля и потребления топлива. Чем тяжелее транспортное средство, тем больший размер двигателя оно имеет и дороже стоит. Также стоит заметить, что расход топлива в городских условиях и расход топлива на трассе имеют сильную корреляцию, равную 0,97. Мощность двигателя также сильно коррелирована с данными величинами, коэффициенты корреляции равны 0,84 и 0,83 соответственно. Другой высоко коррелированной группой являются колёсная база, длина, ширина и собственный вес автомобиля. Колёсная база - это расстояние между центрами переднего и заднего колёс. Колёсная база, длина и ширина являются различными мерами размеров транспортных средств, поэтому логично, что они высоко коррелированы.

Поскольку многие из этих ковариаций сильно коррелируют с друг другом, затея запустить множественную линейную регрессию с использованием всех переменных без учёта проблем мультиколлинеарности будет не очень хорошей идеей.

Ход поршня - это длина хода цилиндра в поршневом двигателе. В поршневом двигателе степень сжатия представляет собой соотношение между объемом цилиндра и камерой сгорания, когда поршень находится в нижней части его хода, и объемом камеры сгорания, когда поршень находится в верхней части его хода. Желательна высокая степень сжатия, поскольку она позволяет двигателю извлекать больше механической энергии из заданной массы воздушно-топливной смеси из-за ее более высокой тепловой эффективности.

2.2 Анализ работ

Каждое исследование не может обойтись без тщательного изучения работ по похожим темам. Данная процедура предназначена прежде всего для осознания актуальности работы в данный момент времени и для повышения компетентности исследователя. Таким образом, перед началом работы встаёт вопрос о проведении литературного обзора.

Обзор литературы представляет собой изучение работ, написанных авторами на тему проводимого исследования. Он необходим, в первую очередь, для описания сделанного по изучаемой теме к началу написания работы: текущее состояние проблемы, предполагаемые различными авторами решения, существующие концепции и т.д..

Для эффективного анализа работ были использованы приведённые ниже ключевые слова: прогноз, регрессионное моделирование, многомерные данные, непараметрическая идентификация.

2.2.1 Обзор научных публикаций по теме бакалаврской работы

Данный параграф посвящён найденным научным публикациям, представляющие интерес для написания бакалаврской работы.

В работе [39] Овечкиной О.О. «Агрегация и регрессионный подход к численному моделированию больших данных» рассматривается обработка, представление и моделирование на основе численных методов выборки данных большого объёма. На основе гистограммного подхода происходит построение процедур агрегации. Используются агрегированные данные для рассмотрения вопросов численного моделирования, необходимого для выявления зависимостей между входными и выходными характеристиками. Автор анализирует ряд подходов построения регрессионных моделей на основе различных метрик в пространстве гистограмм.

В статье [40] представлена информация об электропроводимости осадочного чехла в верхней мантии литосферы, её вещественный состав, геомеханические свойства, термодинамическое состояние, флюидный режим и другие характеристики. Описываются проблемы, которые в настоящий момент стоят перед региональной геоэлектрикой. Первой проблемой является изучение строение нефтегазосодержащих и рудоносных провинций для классификации типов земной коры, которая содержит нефтегазовые бассейны и рудные месторождения. Для решения данной проблемы авторами предпринимается попытка построения модели геоэлектрических зон на основе существующих аналогов. Вторая проблема представляет собой изучение глубинного строения сейсмоактивных зон на основе элементов прогнозирования.

В работе [41] говорится о повышении эффективности планирования водно-энергетических режимов работы Бурейской ГЭС с помощью разработанного метода краткосрочного прогноза притока воды, опирающегося на гидрологическую модель и информацию о метеорологических прогнозов. В качестве модели авторы используют пространственно-распределенную физико-математическую модель формирования стока, построенную на основе

наблюдений гидрометеорологического характера в бассейне реки Бурея. Статистические критерии показали высокое качество моделирования и хороший потенциал дальнейшего прогнозирования. Все расчёты выполняются на основе метеорологических данных, полученных от двух моделей атмосферной циркуляции. Корректировка модели производится с учётом информации о притоке воды в водохранилище. В статье приведены испытания и их результаты о прогнозировании притока воды в Бурейское водохранилище на 2016 год.

2.2.2 Обзор книг и монографий по теме бакалаврской работы

В учебном пособии [42] рассмотрены общепринятые методы обработки многомерных экспериментальных данных объектов различной природы, статистического анализа и представления данных. Также подробно изложены методы многомерной статистической обработки: метод главных компонент, каноническая корреляция, вейвлет-анализ, факторный анализ, дискриминантный анализ, многомерное шкалирование. Описаны современные методы сингулярного разложения, применяемые в обработке многокомпонентных временных рядов. Также в книге имеется множество примеров с иллюстрациями, которые взяты в том числе и на основе многолетней практики решения различных реальных задач самими авторами.

В монографии [43] излагается подход к построению математических моделей стохастических объектов в условиях непараметрической неопределённости, когда не известна структура объекта, а известна только общая информация. Книга содержит не только обширный теоретический материал, но и множество интересных примеров, иллюстраций, приложений, разработанных методов м алгоритмов в различных сферах жизни общества. Данная работа будет интересна специалистам, аспирантам и студентам, занимающимися обработкой информации, прогнозированием и управлением.

2.2.3 Обзор диссертаций по теме бакалаврской работы

Диссертационная работа Корнеевой А.А. [44] посвящена построению непараметрической модели и алгоритмов управления для многомерных систем с запаздыванием. Задача управления сложными промышленными объектами связана с решением задачи идентификации исследуемого объекта. На основе априорной информации выделяют параметрическую и непараметрическую идентификацию. Постановка и решение задачи идентификации напрямую зависят от класса исследуемого объекта. В диссертационной работе проводится исследования нового класса процессов, называемых «трубчатыми» (или H-процессами). Для решения задачи идентификации и управления процессами «трубчатого» типа важную роль играет первичная обработка исходных данных.

Цель работы состоит построении и исследовании непараметрических моделей и алгоритмов управления для многомерных дискретно-непрерывных процессов «трубчатой» структуры с запаздыванием, которые ранее не были исследованы.

Диссертация [45] повествует нам о применимости регрессионного моделирования для решения задач астрометрии и небесной механики. В данной работе показана история развития средств и методов изучения околоземного пространства и дальнего космоса. Для повышения точности астрономических наблюдений необходима адаптация методов математической обработки результатов измерений, например, в расчёте орбит небесных тел. Рассмотрены способы решения задач оценивания. Применение построения моделей на основе регрессии подразумевает исследование и выбор оптимальных методов получения наилучших линейных оценок параметров и тестирование получаемой модели по исходным данным. В данной работе проводится исследование применимости методов параметрической идентификации для решения системы уравнений, описывающих орбитальное движение небесных тел на основе радио-интерферометрических наблюдений внегалактических источников.

Цель диссертационной работы состоит в решении научно-технической задачи разработки, исследования методов оценивания характеристик математических моделей, построенных на теории орбитального движения и вращения планет по лазерным наблюдениям на основе регрессионного моделирования и разработка предметно-ориентированной программной системы.

На рисунке 9 показа тенденция к публикации работ учёными, начиная с 2000 года. Данные для построения графика были взяты с научной электронной библиотеки «eLibrary». Запрос формировался на основе ключевых слов, указанных перед пунктом 2.2.1.

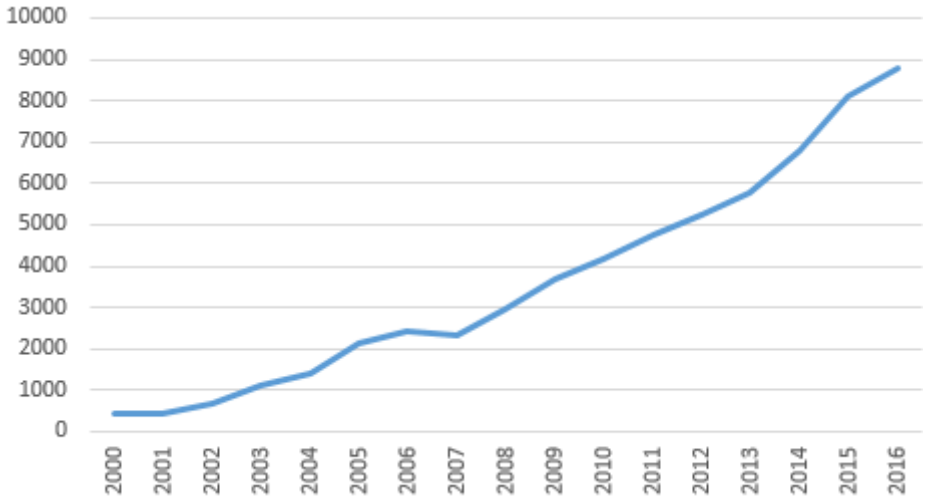

Рисунок 9 – Зависимость публикаций работ от года публикации

Такая тенденция говорит о том, что в последнее время всё больше людей занимаются решением проблемы прогнозирования, и такая статистика не может не радовать.

2.3 Azure

Студия машинного обучения Azure – это служба прогнозной аналитики, позволяющей за короткое время создавать, тестировать и управлять

прогнозными моделями для решения задач аналитики. Исследователю предлагаются готовые библиотеки алгоритмов для быстрого развёртывания прогнозных решений. Базовый рабочий процесс в Azure приведён на рисунке 10:

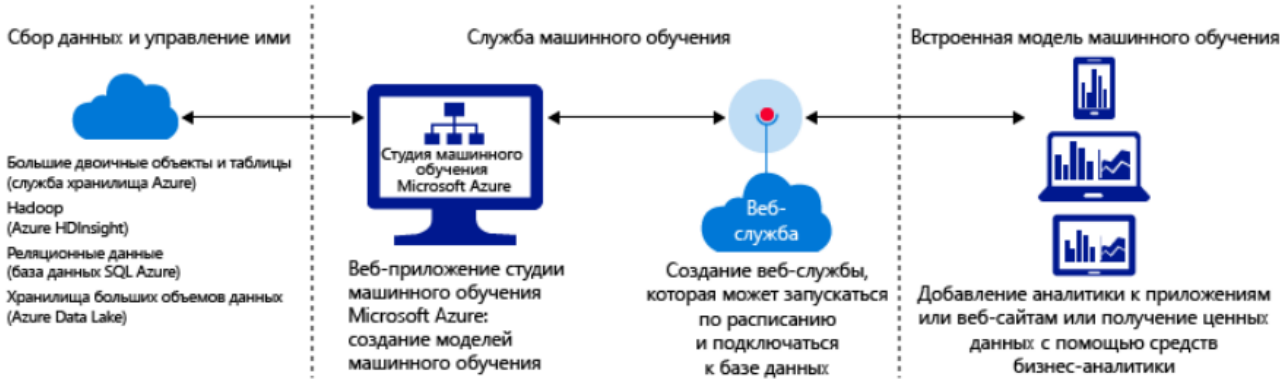

Рисунок 10 – Создание моделей на основе данных и управление решением машинного обучения

Студия машинного обучения Azure позволяет:

- использовать аналитические решения, разработанные другими пользователями или добавлять собственные;

- использовать обширную библиотеку модулей машинного обучения, написанных на языках R и Python, расширять модули собственными сценариями;

- развёртывать предложенные решения в качестве веб-служб.

Для разработки прогнозной модели используют данные от одного или нескольких источников. Для достижения результата данные анализируются и преобразуются различными операциями и вычислениями. Такая разработка модели является итеративным процессом. Параметры различных функций изменяются до тех пор, пока не будет создана эффективная модель.

Azure представляет собой визуальное рабочее пространство, необходимое для создания, тестирования, обучения модели прогнозной аналитики. Исследователь перемещает наборы данных и модули анализа на рабочее пространство и связывает их вместе. Данными манипуляциями он создаёт эксперимент, который будет выполнен в студии машинного обучения. Когда

эксперимент будет готов, происходит его преобразование в прогнозный и исследователь решает, стоит ли опубликовать его как веб-службу, чтобы другие пользователи могли использовать созданный проект в своих целях.

Эксперимент состоит из данных, использующихся в модулях аналитики, соединённых вместе и представляющих модель прогнозной аналитики. Каждый эксперимент имеет следующие характеристики:

- эксперимент состоит минимум из одного набора данных и одного модуля;
- наборы данных связываются только с модулями;
- модули могут быть связаны с другими модулями или с наборами данных;
- все входы модулей обязаны иметь связь с потоком данных;
- должны быть установлены все необходимые параметры модулей.

Набор данных – это информация, загруженная в студию машинного обучения для моделирования необходимой системы.

Модуль – это какой-либо алгоритм для работы с данными. Модули необходимы для ввода данных, оценки и проверки. В ходе построения эксперимента исследователь выбирает из каких модулей будет состоять его проект.

2.3.1 Создание эксперимента на основе существующего решения

Ниже будет представлено создание эксперимента в студии машинного обучения Azure, который будет представлять из себя решение, которое используется в настоящее время для прогнозирования стоимости автомобиля на основе его характеристик.

Каждый эксперимент в студии машинного обучения состоит из пяти шагов:

- получение данных;
- подготовка данных;
- определение признаков;
- выбор и применение алгоритма обучения;

- получение спрогнозированного решения.

Для работы в студии машинного обучения необходимо перейти по ссылке https://studio.azureml.net.

ШАГ 1. Получение данных

Для использования машинного обучения исследователю необходимы данные. Студия машинного обучения содержит некоторое число наборов данных, которые можно использовать для примера. Также имеется возможность импортировать данные из других источников и привести их к необходимому формату. Для загрузки данных перейдём на репозиторий https://archive.ics.uci.edu/ml/datasets/Automobile и загрузим на свой компьютер файл imports-85.data, содержащий нашу выборку по автомобилям.

Создадим новый эксперимент. Для этого выберем **New** (Создать) в нижней части страницы, и выберем **Dataset** и **From Local File**. Откроется меню загрузки, здесь выберем путь до загруженного файла, название и в качестве типа выберем **Generic CSV File with no header (.hn.csv).**

Для создания нового эксперимента последовательно выберем **+NEW** (+Создать), **Experiment** (Эксперимент), а затем **Blank Experiment** (Пустой эксперимент). В верхней части холста можно изменить имя эксперимента. После создания пустого эксперимента оно будет присвоено по умолчанию. Используемое имя не должно быть уникальным! Окно созданного пустого эксперимента приведено на рисунке 11:

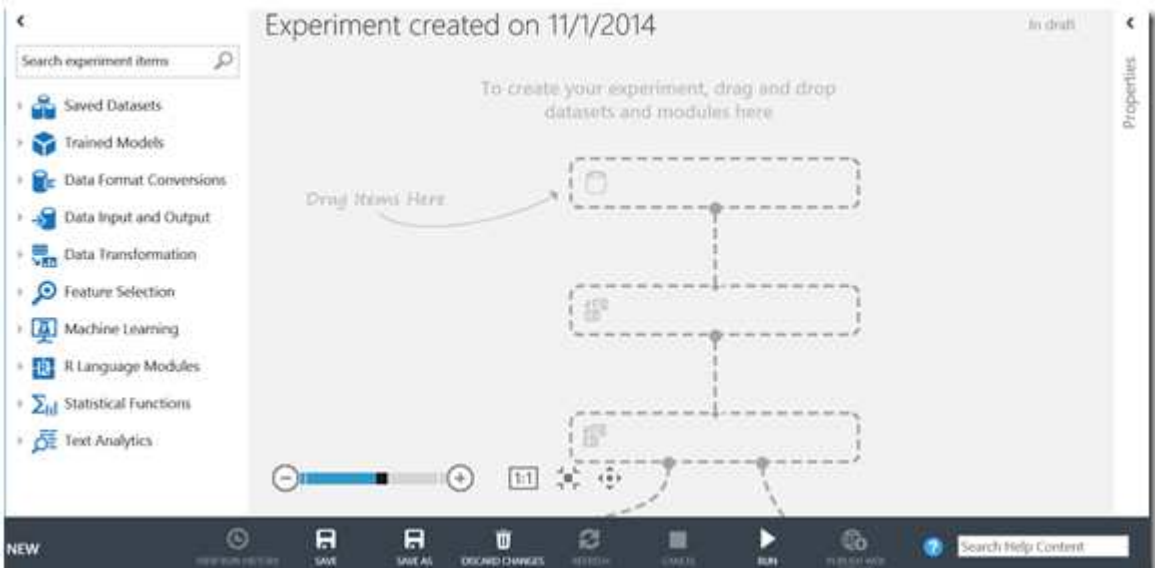

Рисунок 11 – Окно начала эксперимента

Загруженные нами данные находятся в разделе **Saved Datasets** (Сохранённые данные), который расположен слева активного окна эксперимента. Выберем наши данные и перенесём их на любое место рабочего пространства. Для того, чтобы увидеть, как выглядят данные, необходимо щёлкнуть на порт выхода в нижней части набора данных и выбрать пункт ***Visualize*** *(Визуализировать)*.

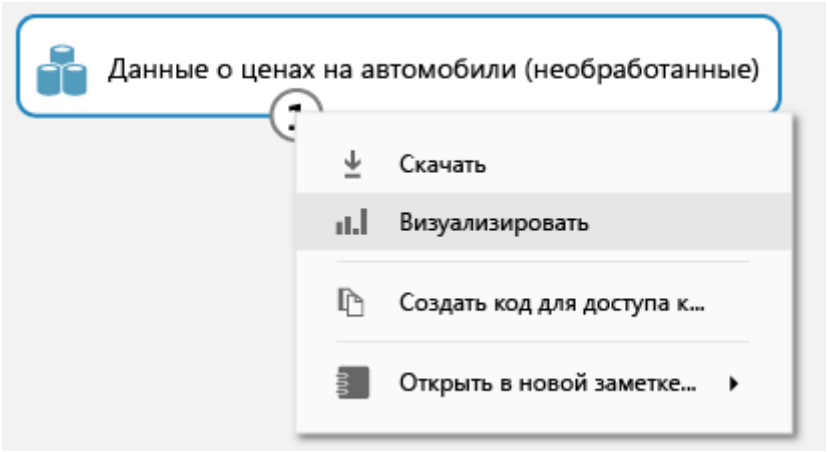

Рисунок 12 – Выбор пункта «Визуализировать» набора данных

Здесь каждый экземпляр автомобиля является строкой, а его характеристики столбцами. Нам необходимо спрогнозировать цену автомобиля

по имеющимся характеристикам и отобразить её в крайнем правом столбце с названием **price** (цена). Закроем окно визуализации.

ШАГ 2. Подготовка данных

Обычно перед анализом необходима предварительная обработка набора данных. Выборка имеет пропуски. Для того, чтобы модель смогла эффективно проанализировать данные, необходимо избавиться от пропущенных значений. В предлагаемом решении нашей проблемы происходит удаление всех строк, содержащих пропуски. Мы поступим также. Столбец «Нормированные потери» также содержит значительное число пропусков, поэтому он тоже будет исключён. Удаление пропусков из выборки является необходимым условием для работы большинства модулей студии машинного обучения Azure.

Во-первых, будет добавлен модуль, который удалит из выборке столбец **normalized-losses** (нормированные потери). Для этого найдём модуль **Select Columns in Dataset** (Выбор столбцов в наборе данных) и перенесём этот модуль на рабочее пространство эксперимента. Теперь необходимо соединить выход набора данных «Данные о ценах на автомобиль (необработанные)» со входом модуля **Select Columns in Dataset**. Затем выберем модуль **Select Columns in Dataset** в области **Свойства** откроем **Launch column selector** (Запустить средство выбора столбцов). В списке **Исключить** выберем элемент **normalized-losses,** чтобы исключить его из выборки. Нажмём кнопку «ОК», чтобы принятые изменения вступили в силу.

Во-вторых, перенесём на холст эксперимента модуль **Clean Missing Data** (Очистка недостающих данных) и соединим его с модулем **Select Columns in Dataset**. Откроем **Properties** (Свойства) и выберем **Remove entire row** (Удалить всю строку) для параметра **Cleaning mode** (Режим очистки). Здесь будут удалятся все строки, которые содержат хотя бы одно пропущенное значение.

Теперь мы можем запустить эксперимент. Кнопка запуска расположена в нижней части страницы. Как только эксперимент будет выполнен рядом с каждым модулем должен появиться зелёный флажок, означающий успешное

выполнение. На рисунке 13 показан приблизительный вид выполнения эксперимента.

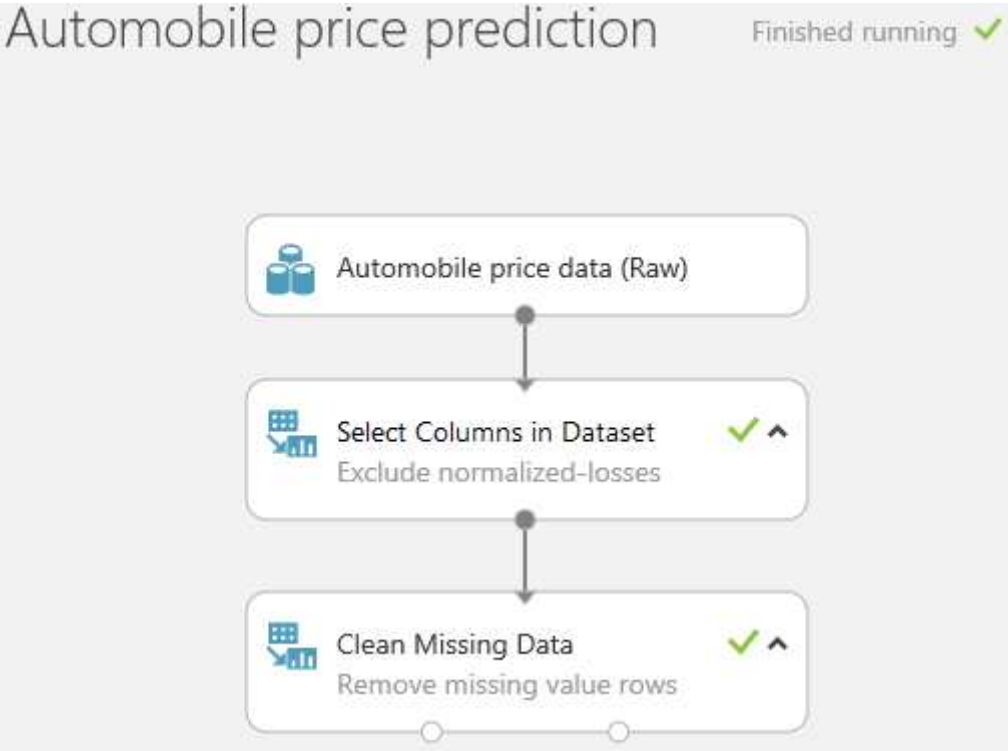

Рисунок 13 – Выполнение эксперимента

Для просмотра очищенного набора данных необходимо выбрать левый выход модуля **Clean Missing Data** и нажать **Visualize** (Визуализировать). Теперь мы готовы задать свойства, которые будет использовать наша прогнозная модель.

ШАГ 3. Определение признаков

В машинном обучении признаками называются отдельные свойства использующихся объектов. В нашем случае каждая строка представляет собой автомобиль, а каждый столбец его признак.

Для определения признаков использовался корреляционный анализ. С его результатами можно познакомиться в пункте 2.1. На основе результатов корреляционного анализа были отобраны следующие признаки: марка, тип кузова, колёсная база, мощность двигателя, размер двигателя, обороты

двигателя, расход топлива на трасе и цена. Первые 7 признаков будут считаться входными, а 8 признак (цена) – выходным, его нам требуется спрогнозировать.

Перенесём на рабочее пространство ещё один модуль **Select Columns in Dataset.** Соединим левый выход модуля **Clean Missing Data** со входом модуля **Select Columns in Dataset.** В области **Свойства** добавленного модуля откроем **Launch column selector** и выберем 8 вышеперечисленных признаков в фильтре Включить. Нажмём кнопку «ОК».

В результате мы получим набор данных, содержащий только те признаки, которые будут переданы в обучающий алгоритм.

ШАГ 4. Выбор и применение алгоритма обучения.

Теперь полученные данные будут использоваться для обучения модели и её тестирования. Имеется два типа алгоритмов машинного обучения, на основе которых можно сделать прогноз – это классификация и регрессия. Первая используется для формирования прогноза по заданному набору категорий. Вторая применяется для прогнозирования числа. Нам необходимо спрогнозировать цену, а это число, поэтому будет использоваться модель регрессии. В существующем решении используется линейная регрессия. Следовательно мы также на основе линейной регрессии будем делать прогноз.

При обучении модели мы должны передать ей выборку, содержащую цены. Модель ищет зависимости между признаками автомобиля и его ценой. Далее происходит тестирование модели – происходит проверка точности спрогнозированной цены и реальной.

Полученную выборку мы разделим на две части, первая будет применятся для обучения модели, вторая для её тестирования.

Перенесём на рабочее пространство модуль **Split Data** (Разделение данных) и соединим его с выходом модуля **Select Columns in Dataset**. Выберем модуль **Split Data.** На панели свойства найдём параметр **Fraction of rows in the first output dataset** (Доля строк в первом выходном наборе данных) и установим для него значение 0,75. Это показывает нам, что 75% выборки будет

использовано для обучения модели, а 25% - для тестирования. Также здесь можно выбрать распределение выборки (по умолчанию стоит случайное).

Запустим эксперимент, чтобы добавленные модули смогли передать определения столбцов, которые будут добавлены нами позднее.

В качестве алгоритма обучения будет использоваться линейная регрессия. Выберем модуль **Linear Regression** (Линейная регрессия). Также добавим на холст модуль **Train Model** (Обучение модели), затем соединим выход модуля **Linear Regression** с левым входом модуля **Train Model**, а левый выход модуля **Split Data** с правым входом модуля **Train Model.** Далее в свойствах модуля **Train Model** необходимо выбрать столбец, который наша модель должна спрогнозировать, то есть столбец **цена.**

Запустим эксперимент. Теперь мы имеем обученную регрессионную модель, которая будет использована для оценки новых данных об автомобилях для прогнозирования их цены. После выполнения эксперимент он должен выглядеть примерно так, как показано на рисунке 14.

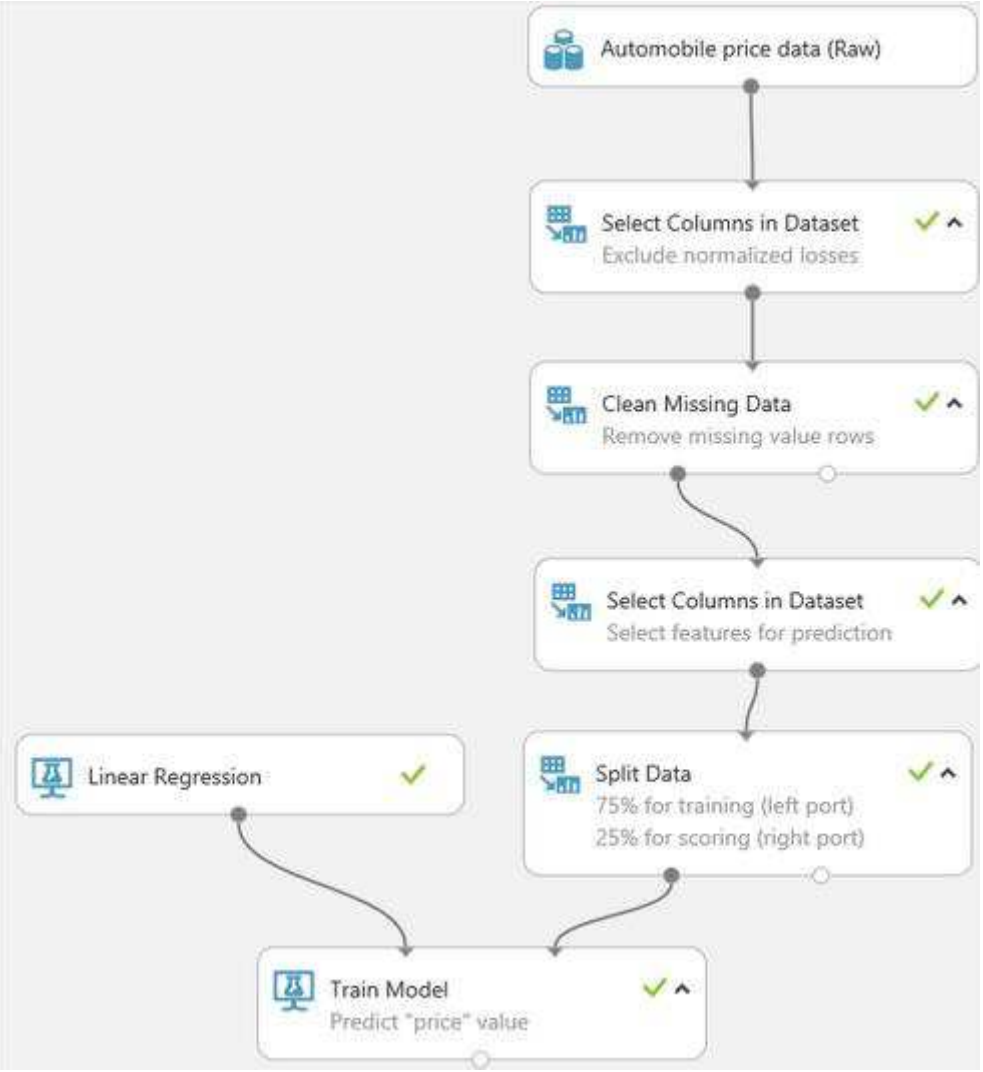

Рисунок 14 – Приблизительный вид эксперимента на ШАГе 4

ШАГ 5. Получение спрогнозированного решения

После обучения модели на основе 75% данных, она будет использоваться для оценки оставшихся 25% данных.

Перенесём на рабочее пространство эксперимента модуль **Score Model** (Оценка модели). Соединим его левый вход с выходом модуля **Train Model,** а правый вход с правым выходом модуля **Split Data**. Запустим эксперимент. Для проверки выходных данных модуля **Score Model** необходимо выбрать пункт **Визуализировать**. После выполнения этой операции будут показаны прогнозируемые значения цены вместе с известными значениями имеющихся данных. Это можно наблюдать на рисунке 15.

Рисунок 15 – Выходные данные модуля «Оценка модели»

Теперь мы можем проверить качество результатов. Добавим модуль **Evaluate Model** (Анализ модели) и соединим его с выходом модуля **Score Model**. Запустим эксперимент. Для проверки выходных данных модуля **Evaluate Model,** выберем элемент **Визуализировать**. Результат работы модели прогнозирования представлен на рисунке 16.

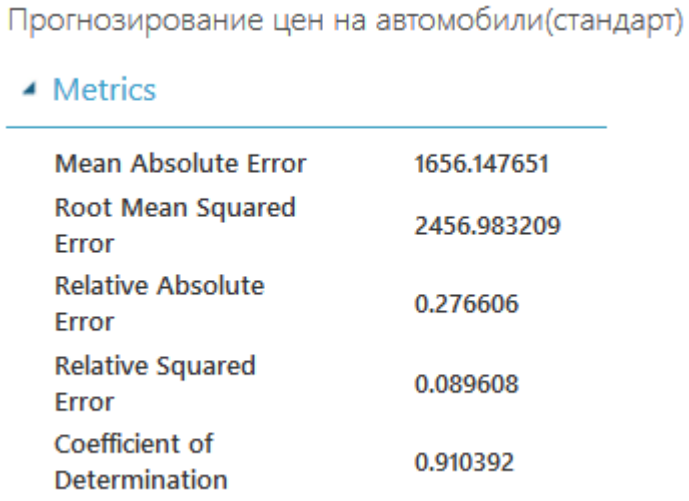

Рисунок 16 – Оценка результатов эксперимента

Здесь:

- **Mean Absolute Error** (Средняя абсолютная ошибка) – среднее значение арифметических отклонений;

- **Root Mean Squared Error** (Средняя среднеквадратическая ошибка) - квадратный корень из среднего значения возведенных в квадрат арифметических отклонений спрогнозированных значений тестового набора данных;

- **Relative Absolute Error** (Относительная абсолютная ошибка) - среднее арифметическое отклонение по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений;

- **Relative Squared Error** (Относительная квадратичная ошибка) - среднее арифметическое среднеквадратичных отклонений по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений;

- **Coefficient of Determination** (Коэффициент детерминации) - статистический показатель, который оценивает соответствие модели данным.

Окончательный вид эксперимента представлен на рисунке 17.

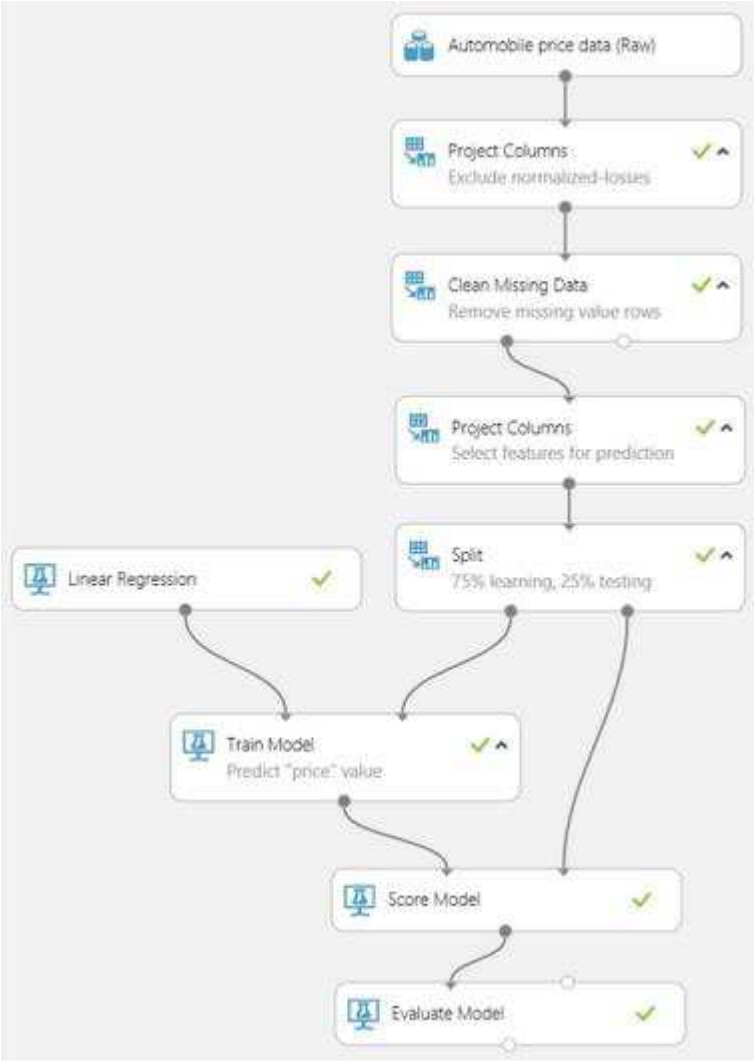

Рисунок 17 – Итоговый вид эксперимента

2.4 Предлагаемое решение

Студия машинного обучения Azure предлагает только параметрический подход к решению задачи идентификации. Для повышения точности прогноза в условиях малого объёма данных и недостатка априорной информации лучше будет использовать непараметрический подход.

Для решения задачи непараметрического восстановления регрессии в работе применяется оценка Надарая-Ватсона. Эта оценка была предложена Е. Надарая и Дж. Ватсоном в 1964 году. Она приведена в формуле (7) и описана в работе [22]. Для большей наглядности продублируем её:

$$x_s(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}, \quad (7)$$

где $u = (u_1, u_2, ..., u_m)$ - m-мерный вектор входных воздействий;

x - выходная переменная;

$\Phi(c_s^{-1}(u-u_i))$ - ядерная колокообразная функция;

c_s - коэффициент размытости ядра.

В качестве ядерной колокообразной функции будут применятся функции, содержащие треугольное и параболическое ядро. На основе эксперимента будет выбрано ядро, наиболее подходящее для решения задачи восстановления регрессии. Коэффициент размытости ядра будет лежать в пределах [0,1;1,5] и его настройка будет зависеть от решения задачи минимизации квадратичного показателя, опирающегося на метод скользящего экзамена, приведённого в формуле (12).

$$R(c_s) = \sum_{k=1}^{s} (x_k - x_s(u_k, c_s))^2 = \min, k \neq i. \quad (12)$$

Для обучения непараметрической модели исходная выборка будет разделена на две части: обучающую и экзаменационную. Формула Надарая-Ватсона (7) будет немного изменена, чтобы лучше решать задачу прогнозирования. Полученная формула будет иметь вид, представленный в формуле (0):

$$y_s(u) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}, \quad (18)$$

где $u = (u_1, u_2, ..., u_m)$ - m-мерный вектор входных воздействий;

y_s – спрогнозированная цена;

y_i – цена i-го наблюдения обучающей выборки;

s – объём обучающей выборки;

u^j – j-ое входное воздействие прогнозируемого наблюдения;

u_i^j - j-ое входное воздействие i-го наблюдения обучающей выборки;

$\Phi(c_s^{-1}(u-u_i))$ - ядерная колокообразная функция;

c_s - коэффициент размытости ядра.

Выводы по 2 главе

На основании написанного во второй главе можно сделать следующие выводы:

1) были проанализированы и описаны используемые данные;
2) анализ существующих работ по заданной тематике показал тенденцию к популяризации решения задач прогнозирования;
3) рассмотрено существующие решение прогнозирования стоимости автомобиля;
4) реализован параметрический алгоритм прогнозирования в студии машинного обучения Azure;
5) предложено решение непараметрического подхода к восстановлению регрессии.

С помощью проведённых исследований будут выбраны методы проведения эксперимента и получен непараметрический алгоритм прогнозирования стоимости автомобиля.

3 Разработка непараметрического алгоритма прогнозирования стоимости автомобиля

3.1 Решение задачи заполнения пропусков в матрице наблюдений

Для того, чтобы решить задачу заполнения пропусков в матрице наблюдений для начала необходимо проанализировать, как ведут себя различные методы импутирования на обучающей выборки максимально приближенной к экзаменационной.

В качестве обучающей выборки было взято 40% наблюдений из исходной матрицы наблюдений, не содержащих пропуски ни в одной переменной. Также пришлось удалить признак «Нормированные потери», так как около 70% пропущенных значений в исходной выборке находились именно в нём, и не один метод заполнения не сможет эффективно работать с таким количеством пропусков.

С помощью генератора случайных чисел в обучающей выборке генерировались пропуски, количество которых соответствовало 5%, 10%, 15% значений от числа элементов обучающей выборки. Это делалось для того, чтобы оценить потерю информации в результате импутирования. Так как в исходной матрице в категориальных переменных количество пропусков меньше 3% от совокупности всех пропущенных значений, то было принято решение не генерировать их в категориальных переменных.

Для сравнения точности импутирования пропущенные значения заполнялись тремя методами: методом Бартлета, Resampling-методом и линейным регрессионным моделированием. Достоинства и недостатки каждого из них описаны в параграфе 1.3. Для проведения эксперимента использовалось приложение SPSS Statistica 17.0.

Статистика сгенерированных пропусков приведена в таблице 4.

Таблица 4 – Количество сгенерированных пропусков для каждой переменной

№	НАЗВАНИЕ ПРИЗНАКА	ЧИСЛО СГЕНЕРИРОВАННЫХ ПРОПУСКОВ		
		5%	10%	15%
1	Рейтинг	2	6	8
2	Колёсная база	4	5	12
3	Длина	6	9	11
4	Ширина	2	11	13
5	Высота	3	6	13
6	Масса	5	4	12
7	Размер двигателя	4	7	9
8	Диаметр выхлопной трубы	6	7	12
9	Ход поршня	3	6	7
10	Коэффициент компрессии	2	6	10
11	Мощность двигателя	1	9	11
12	Обороты двигателя	4	8	5
13	Расход топлива в городе	5	7	13
14	Расход топлива на трассе	2	9	10
15	Цена	3	6	11
ИТОГО		52	106	157

Для оценивания точности каждого метода импутирования рассчитывалась среднеквадратическая ошибка, её вид показан в формуле (19).

$$W = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x^i - x_s^i)^2}, \tag{19}$$

где n – количество наблюдений;

x^i – реальное значение;

x^i_s – импутированное значение.

Минимизация ошибки (19) позволит выбрать метод заполнения, наиболее подходящий для заполнения пропусков в исходной матрице наблюдений. Для

адекватного значения ошибки перед её расчётом все параметры были приведены к одной степени точности.

Результаты проведённого эксперимента приведены в таблице 5:

Таблица 5 – Результаты эксперимента по оцениванию точности методов импутирования

МЕТОД	**СРЕДНЕКВАДРАТИЧЕСКАЯ ОШИБКА**		
	5%	**10%**	**15%**
Бартлета	0,36	0,3	0,39
Resampling	0,3	0,35	0,73
Линейная регрессия	0,23	0,28	0,4

Из результатов проведённого эксперимента видно, что на метод Бартлета почти не оказывает влияния количество пропусков в матрице наблюдений, Resampling-метод показывает себя не удовлетворительно с повышением количества пропусков, а импутирование на основе линейной регрессии немного ухудшается с повышением количества пропусков.

Таким образом, по результатам эксперимента, а также не забывая тот факт, что количество пропусков в исходной матрице наблюдений составляет 8,78% от числа всех наблюдений, было принято решение в качестве метода заполнения пропусков использовать на основе линейного регрессионного моделирования.

3.2 Выявление значимых признаков

После того, как были заполнены пропуски в матрице наблюдений, самое время перейти к определению информативных признаков и сокращению размерности массива данных.

Как уже говорилось выше, на сегодняшний день для решения задачи выделения информативных признаков из набора данных, используемого в работе, применяют корреляционный анализ. Но корреляционный анализ не является универсальным методом сокращения размерности, поэтому вместе с ним будет также рассмотрен метод главных компонент, описанный в параграфе 1.4 и реализованный в программном обеспечении SPSS Statistica 17.0.

3.2.1 Корреляционный анализ

Корреляционный анализ будет выполнен на основе мер близости между признаками. Коэффициент корреляции рассчитывается по формуле (17) и для более наглядного представления пересчитывается для предела [0,1]. Формула (17) продублирована ниже:

$$r_{x,y} = \frac{1}{s}\sum \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y}, \quad (17)$$

где, s - объём выборки;

x,y - признаки;

m_x, m_y - математические ожидания признаков x и y;

σ_x, σ_y - среднеквадратические отклонения признаков x и y.

Признаки, используемые в анализе, расписаны в параграфе 2.1. Используется 24 входные переменные x1, x3, x4,…,x25 (признак x2 – «Нормированные потери» не используется в анализе, причины приведены в параграфе 3.1) и одна выходная переменная y (цена).

Результаты проведённого корреляционного анализа приведены в таблице 6.

Таблица 6– Корреляционная матрица признаков

ПРИЗНАК	X1	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
X1	1,000	0,384	0,561	0,417	0,076	0,676	0,474	0,331	0,356	0,489	0,530	0,477	0,322
X3	0,384	1,000	0,387	0,482	0,537	0,363	0,453	0,420	0,414	0,375	0,453	0,511	0,464
X4	0,561	0,387	1,000	0,224	0,344	0,540	0,526	0,428	0,419	0,474	0,501	0,478	0,328
X5	0,417	0,482	0,224	1,000	0,481	0,439	0,413	0,483	0,524	0,420	0,407	0,414	0,635
X6	0,076	0,537	0,344	0,481	1,000	0,196	0,510	0,530	0,505	0,450	0,292	0,396	0,559
X7	0,676	0,363	0,540	0,439	0,196	1,000	0,483	0,522	0,367	0,499	0,446	0,401	0,315
X8	0,474	0,453	0,526	0,413	0,510	0,483	1,000	0,534	0,387	0,424	0,367	0,426	0,126
X9	0,331	0,420	0,428	0,483	0,530	0,522	0,534	1,000	0,472	0,474	0,206	0,475	0,422
X10	0,356	0,414	0,419	0,524	0,505	0,367	0,387	0,472	1,000	0,522	0,458	0,510	0,650
X11	0,489	0,375	0,474	0,420	0,450	0,499	0,424	0,474	0,522	1,000	0,492	0,389	0,442
X12	0,530	0,453	0,501	0,407	0,292	0,446	0,367	0,206	0,458	0,492	1,000	0,432	0,417
X13	0,477	0,511	0,478	0,414	0,396	0,401	0,426	0,475	0,510	0,389	0,432	1,000	0,421
X14	0,322	0,464	0,328	0,635	0,559	0,315	0,126	0,422	0,650	0,442	0,417	0,421	1,000
X15	0,479	0,422	0,498	0,393	0,415	0,433	0,517	0,386	0,450	0,599	0,442	0,403	0,420
X16	0,387	0,380	0,465	0,424	0,440	0,391	0,273	0,347	0,611	0,384	0,501	0,428	0,796
X17	0,391	0,411	0,412	0,512	0,459	0,353	0,154	0,340	0,662	0,418	0,485	0,437	0,932
X18	0,525	0,521	0,558	0,576	0,427	0,436	0,220	0,389	0,533	0,448	0,433	0,458	0,772
X19	0,440	0,472	0,383	0,508	0,443	0,469	0,584	0,473	0,426	0,381	0,390	0,499	0,306
X20	0,495	0,606	0,513	0,410	0,431	0,497	0,533	0,478	0,395	0,411	0,565	0,443	0,359
X21	0,426	0,506	0,160	0,586	0,536	0,417	0,439	0,496	0,445	0,455	0,408	0,410	0,533
X22	0,488	0,429	0,546	0,577	0,383	0,417	0,175	0,278	0,585	0,463	0,523	0,443	0,863
X23	0,605	0,330	0,721	0,345	0,316	0,531	0,476	0,340	0,374	0,495	0,613	0,427	0,301
X24	0,431	0,481	0,306	0,336	0,443	0,505	0,705	0,538	0,322	0,387	0,431	0,513	0,023
X25	0,470	0,479	0,343	0,307	0,430	0,523	0,707	0,509	0,296	0,390	0,417	0,510	0,000
Y	0,404	0,360	0,390	0,555	0,469	0,357	0,127	0,265	0,651	0,487	0,528	0,400	0,923

Продолжение таблицы 6

ПРИЗНАК	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	Y
X1	0,479	0,387	0,391	0,525	0,440	0,495	0,426	0,488	0,605	0,431	0,470	0,404
X3	0,422	0,380	0,411	0,521	0,472	0,606	0,506	0,429	0,330	0,481	0,479	0,360
X4	0,498	0,465	0,412	0,558	0,383	0,513	0,160	0,546	0,721	0,306	0,343	0,390
X5	0,393	0,424	0,512	0,576	0,508	0,410	0,586	0,577	0,345	0,336	0,307	0,555
X6	0,415	0,440	0,459	0,427	0,443	0,431	0,536	0,383	0,316	0,443	0,430	0,469
X7	0,433	0,391	0,353	0,436	0,469	0,497	0,417	0,417	0,531	0,505	0,523	0,357
X8	0,517	0,273	0,154	0,220	0,584	0,533	0,439	0,175	0,476	0,705	0,707	0,127
X9	0,386	0,347	0,340	0,389	0,473	0,478	0,496	0,278	0,340	0,538	0,509	0,265
X10	0,450	0,611	0,662	0,533	0,426	0,395	0,445	0,585	0,374	0,322	0,296	0,651
X11	0,599	0,384	0,418	0,448	0,381	0,411	0,455	0,463	0,495	0,387	0,390	0,487
X12	0,442	0,501	0,485	0,433	0,390	0,565	0,408	0,523	0,613	0,431	0,417	0,528
X13	0,403	0,428	0,437	0,458	0,499	0,443	0,410	0,443	0,427	0,513	0,510	0,400
X14	0,420	0,796	0,932	0,772	0,306	0,359	0,533	0,863	0,301	0,023	0,000	0,923
X15	1,000	0,494	0,474	0,418	0,448	0,327	0,423	0,456	0,454	0,403	0,406	0,484
X16	0,494	1,000	0,929	0,627	0,396	0,372	0,480	0,826	0,379	0,199	0,187	0,851
X17	0,474	0,929	1,000	0,726	0,302	0,325	0,483	0,897	0,314	0,081	0,068	0,944
X18	0,418	0,627	0,726	1,000	0,332	0,378	0,393	0,832	0,510	0,050	0,071	0,736
X19	0,448	0,396	0,302	0,332	1,000	0,540	0,621	0,293	0,405	0,637	0,632	0,326
X20	0,327	0,372	0,325	0,378	0,540	1,000	0,509	0,359	0,466	0,543	0,550	0,359
X21	0,423	0,480	0,483	0,393	0,621	0,509	1,000	0,410	0,246	0,508	0,501	0,483
X22	0,456	0,826	0,897	0,832	0,293	0,359	0,410	1,000	0,533	0,006	0,028	0,879
X23	0,454	0,379	0,314	0,510	0,405	0,466	0,246	0,533	1,000	0,385	0,420	0,397
X24	0,403	0,199	0,081	0,050	0,637	0,543	0,508	0,006	0,385	1,000	1,000	0,063
X25	0,406	0,187	0,068	0,071	0,632	0,550	0,501	0,028	0,420	1,000	1,000	0,053
Y	0,484	0,851	0,944	0,736	0,326	0,359	0,483	0,879	0,397	0,063	0,053	1,000

Возьмём порог значимости равным 0,6. Тогда все переменные, для которых коэффициент корреляции будет больше порога значимости имеют между собой зависимость. В таблице 6 они подсвечены для наглядности. Из таких признаков остаётся в выборке только один признак – остальные удаляются.

Так как цена (Y) является выходной величиной она обязательно должна остаться в выборке наблюдений. С ней сильно коррелируют такие входные величины, как X10, X14, X16, X18, следовательно они удаляются из выборки. Коэффициент корреляции переменных X24 и X25 максимален, также эти они сильно коррелируют с переменными X8 и X19. Оставим переменную X25, а остальные удалим. Также оставив такие признаки, как X3, X7, X10, X17 X22, X23.

Таким образом остались 8 переменных: марка, тип кузова, колёсная база, мощность двигателя, размер двигателя, обороты двигателя, расход топлива на трасе и цена. Именно такие переменные используют в существующем решении задачи прогнозирования на исходных данных. Корреляционный анализ не может дать точного ответа, какие признаки стоит оставлять, а какие удалять, он может выдать огромное количество решений, на проверку которых уйдёт большое количество времени, а результат с максимальной точностью решения может быть и не достигнут. Поэтому было принято решение использовать метод главных компонент в качестве решения задачи сокращения размерности.

3.2.2 Метод главных компонент

Данный метод более подробно описан в главе 1.4. Программная среда SPSS Statistica позволяет реализовать его в пару щелчков мыши. Метод главных компонент основывается на замене имеющихся признаков, так называемыми, главными компонентами. Каждая такая компонента представляет собой некоторую модель, в которой каждому признаку присвоен свой определённый коэффициент.

В качестве входных переменных анализа применяются переменные x1, x3,…, x25. В качестве переменной анализа используется переменная y. Матрица полученных компонент приведена в таблице 7.

Таблица 7 – Матрица коэффициентов оценок компонент

	КОМПОНЕНТА								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
X1	-0,077	0,636	0,476	0,250	0,053	0,033	0,179	0,025	-0,045
X3	-0,045	-0,255	0,184	-0,506	0,247	-0,262	0,391	0,117	0,022
X4	0,047	0,685	-0,399	-0,189	0,244	0,116	-0,003	0,325	0,059
X5	0,239	-0,353	0,367	0,269	0,262	-0,303	0,151	-0,132	-0,133
X6	0,032	-0,641	-0,509	-0,223	0,054	-0,233	-0,093	0,058	-0,053
X7	-0,185	0,428	0,319	0,454	0,234	0,109	-0,216	0,080	0,126
X8	-0,610	0,000	-0,204	0,049	0,034	-0,025	0,071	0,415	-0,083
X9	-0,216	-0,314	-0,240	0,369	0,521	0,192	-0,155	0,045	0,288
X10	0,370	-0,199	-0,132	0,063	-0,128	0,225	0,161	0,020	0,552
X11	0,037	0,154	-0,237	0,416	-0,206	-0,430	0,195	-0,073	0,524
X12	0,077	0,395	0,282	-0,400	-0,442	-0,253	-0,104	-0,171	0,251
X13	-0,057	-0,006	0,071	-0,236	0,133	0,554	0,588	-0,228	0,184
X14	0,904	-0,311	0,061	0,025	0,053	-0,008	-0,058	-0,014	0,041
X15	0,033	0,136	-0,280	0,318	-0,458	-0,159	0,444	0,347	-0,138
X16	0,695	-0,052	0,056	-0,065	-0,323	0,370	-0,215	0,309	-0,008
X17	0,892	-0,127	0,076	-0,011	-0,207	0,239	-0,102	0,090	0,024
X18	0,716	0,164	0,068	-0,027	0,399	-0,102	0,191	0,039	-0,110
X19	-0,357	-0,221	0,321	0,034	-0,039	0,132	0,162	0,490	-0,064
X20	-0,251	0,048	0,312	-0,438	0,234	-0,204	-0,199	0,366	0,443
X21	-0,035	-0,536	0,487	0,191	-0,151	-0,169	-0,038	0,200	0,058
X22	0,893	0,194	0,109	-0,046	0,018	0,019	0,013	0,082	-0,059
X23	-0,047	0,708	-0,155	-0,131	0,069	-0,164	-0,072	-0,046	-0,084
X24	-0,886	-0,175	0,102	-0,055	-0,196	0,187	-0,049	-0,110	0,020
X25	-0,893	-0,097	0,104	-0,063	-0,175	0,177	-0,038	-0,084	-0,020

В таблице 8 приведена статистика общностей. Она показывает какая часть каждого признака была использована при составлении оценки коэффициентов.

Таблица 8 – Таблица общностей

ПРИЗНАК	ОБЩНОСТИ	
	НАЧАЛЬНЫЕ	ИЗВЛЕЧЁННЫЕ
X1	1,000	0,737
X3	1,000	0,654
X4	1,000	0,848
X5	1,000	0,607
X6	1,000	0,793
X7	1,000	0,661
X8	1,000	0,602
X9	1,000	0,756
X10	1,000	0,596
X11	1,000	0,800
X12	1,000	0,764
X13	1,000	0,821
X14	1,000	0,927
X15	1,000	0,770
X16	1,000	0,876
X17	1,000	0,936
X18	1,000	0,764
X19	1,000	0,569
X20	1,000	0,821
X21	1,000	0,659
X22	1,000	0,861
X23	1,000	0,591
X24	1,000	0,917
X25	1,000	0,893

Ковариационная матрица главных компонент приведена в таблице 9.

Таблица 9 – Ковариационная матрица оценок компонент

КОМПОНЕНТА	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	1

Таким образов в результате использования метода главных компонент размерность исходной матрицы уменьшилась до 9 входных признаков и 1 выходного. При этом было потеряно минимум данных.

3.3 Решение задачи прогноза

Для решении задачи прогноза была разработана программа, написанная на языке C# и основанная на составлении прогноза, опираясь на непараметрическую оценку Надарая-Ватсона, приведённую в формуле (18). Продублируем её ещё раз.

$$y_s(u) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{c_s}\right)}, \qquad (18)$$

где $u = (u_1, u_2, ..., u_m)$ - m-мерный вектор входных воздействий;

y_s – спрогнозированная цена;

y_i – цена i-го наблюдения обучающей выборки;

s – объём обучающей выборки;

u^j – j-ое входное воздействие прогнозируемого наблюдения;

u_i^j - j-ое входное воздействие i-го наблюдения обучающей выборки;

$\Phi(c_s^{-1}(u-u_i))$ - ядерная колокообразная функция;

c_s - коэффициент размытости ядра.

Первоначальная выборка, состоящая из 205 наблюдений, разбивается на две части: обучающую выборку и экзаменационную. В состав обучающей выборки входит 75% наблюдений первоначальной выборки, а в состав экзаменационной соответственно 25%. Наблюдения будут отобраны с помощью генератора случайных чисел. Значения признаков приводятся к одному порядку. В оценке (18) будут тестироваться два типа ядерной функции: треугольное ядро и параболическое. На основе оценки (18) происходит настройка коэффициента размытости ядра, который лежит в пределах [0,1;1,5]. После его настройки происходит расчёт стоимости автомобилей, по данным взятым из

экзаменационной выборки. Следующим шагом рассчитывается среднеквадратическая ошибка, показанная в формуле (19):

$$W = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x^i - x_s^i)^2}, \tag{19}$$

где n – объём экзаменационной выборки;

x^i – реальное значение цены;

x^i_s – спрогнозированное значение цены.

Эксперимент был проведён на нескольких видов данных:

- первоначального набора данных с удалением строк, содержащих пропуски (НД1);

- первоначального набора данных с импутированными значениями на месте пропусков методом, описанном в параграфе 3.1 (НД2);

- набора данных с заполненными пропусками и отбором информативных признаков на основе корреляционного анализа (НД3);

- набора данных с удалением строк, содержащих пропуски, и отбором информативных признаков на основе корреляционного анализа (НД4);

- набора импутированных данных сокращённой размерности на основе метода главных компонент (НД5).

Результаты эксперимента приведены в таблице 10.

Таблица 10 – Результаты непараметрического эксперимента

	НАБОРЫ ДАННЫХ									
	НД1		**НД2**		**НД3**		**НД4**		**НД5**	
Ядра	Δ	∩	Δ	∩	Δ	∩	Δ	∩	Δ	∩
Объём экзаменационной выборки	48		51		51		48		51	
Объём обучающей выборки	145		154		154		145		154	
Количество информативных признаков	24		24		7		7		9	
Оптимальный c_s	1,5	1,5	1,5	1,3	0,6	0,7	1,0	1,4	1,5	0,4
Среднеквадратическая ошибка	1825,8	1779,5	1712,3	1672,88	8865	9590,3	8157,1	8079,9	4218,2	4188,7

На основании эксперимента можно сделать следующие выводы:

1) параболическое ядро показывает лучший результат, чем треугольное ядро;
2) среднеквадратичная ошибка меньше у опытов, для которых производилась импутация пропущенных значений, чем у опытов, у которых просто удалялись наблюдения с пропусками;
3) опыты, у которых данные подверглись работе методов сокращения размерности, имеют ошибку, в несколько раз превышающую своё значение, если сравнить с опытами, для которых использовались все имеющиеся параметры. Это стало небольшой неожиданностью. Но объяснить такую тенденцию можно, используя теорию идентификации: непараметрическому подходу не так важно наличие априорной информации, как для параметрического подхода. Поэтому для работы в условиях непараметрической неопределённости и в условии недостатка априорной информации важно использовать всю имеющуюся апостериорную информацию;
4) лучший результат показывает набор данных, для которого не проводились никакие манипуляции с признаками, и у которого использовался метод импутирования пропущенных значений.

В следующем параграфе будет произведено сравнение с существующими способами решения задачи прогнозирования стоимости автомобиля с полученным непараметрическим алгоритмом.

3.4 Сравнение предложенного метода с существующими

После того, как был предложен непараметрический алгоритм прогнозирования стоимости автомобиля, самое время произвести его сравнительный анализ с существующими решениями проблемы прогнозирования. Единственное решение, доступное в свободном доступе,

подробно описано в параграфе 2.3 так же, как и реализация этого решения в студии машинного обучения Azure.

Данное решение включает в себя удаление строк, содержащих пропуски, использование корреляционного анализа в качестве метода сокращения размерности матрицы наблюдений, выбор линейной регрессии, как средства настройки параметрической модели алгоритма, тестирование решения и расчёт ошибки прогнозирования.

Эксперимент для сравнительного анализа параметрического и непараметрического подходов будет основываться на тех же пяти видов набором данных, которые были описаны в предыдущем параграфе 3.3. Продублируем их:

- первоначальный набор данных с удалением строк, содержащих пропуски (НД1);

- первоначальный набор данных с импутированными значениями на месте пропусков методом, описанном в параграфе 3.1 (НД2);

- набор данных с заполненными пропусками и отбором информативных признаков на основе корреляционного анализа (НД3);

- набор данных с удалением строк, содержащих пропуски, и отбором информативных признаков на основе корреляционного анализа (НД4);

- набор импутированных данных сокращённой размерности на основе метода главных компонент (НД5).

НД4 используется в существующем решении задачи прогнозирования стоимости автомобиля.

Эксперимент будет проведён в студии машинного обучения Azure. В качестве алгоритма обучения модели будет использована линейная регрессия. Для получения результатов эксперимента максимально приближенным к результатам непараметрического эксперимента исходная выборка наблюдений будет разделена на обучающую и экзаменационную в таких же пропорциях, как и в эксперименте 3.3, то есть 75% и 25% наблюдений соответственно. За попадание наблюдений в одну из выборок будет отвечать генератор случайных

чисел. За оценивание работы алгоритма также будет отвечать расчёт среднеквадратической ошибки (19):

$$W = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x^i - x_s^i)^2}, \tag{19}$$

где n – объём экзаменационной выборки;

x^i – реальное значение цены;

x^i_s – спрогнозированное значение цены.

Результаты проведённого эксперимента приведены в таблице 11. Также для более наглядного сравнения существующего и предложенного решений результаты непараметрического эксперимента будут приведены ниже.

Таблица 11 – Результаты параметрического эксперимента

	НАБОРЫ ДАННЫХ				
	НД1	**НД2**	**НД3**	**НД4**	**НД5**
Объём экзаменационной выборки	48	51	51	48	51
Объём обучающей выборки	145	154	154	145	154
Количество информативных признаков	24	24	7	7	9
Среднеквадратическая ошибка	2560,75	2488,23	2606,54	2456,98	4641,42

Таблица 10 – Результаты непараметрического эксперимента

	НАБОРЫ ДАННЫХ									
	НД1		**НД2**		**НД3**		**НД4**		**НД5**	
Ядра	Δ	∪	Δ	∪	Δ	∪	Δ	∪	Δ	∪
Объём экзаменационной выборки	48		51		51		48		51	
Объём обучающей выборки	145		154		154		145		154	
Количество информативных признаков	24		24		7		7		9	
Оптимальный c_s	1,5	1,5	1,5	1,3	0,6	0,7	1,0	1,4	1,5	0,4
Среднеквадратическая ошибка	1825,8	1779,5	1712,3	1672,88	8865	9590,3	8157,1	8079,9	4218,2	4188,7

Из пяти проведённых опытов наилучший результат показывает существующий метод решения задач прогнозирования. Даже при импутировании пропусков и использовании тех же признаков результат получается хуже, чем при удалении наблюдений с пропусками. Это может быть вызвано малым количеством априорной информации, которой не хватает для построения более точной параметрической модели. Некоторые результаты прогнозирования, основанного на данных, отобранных методом главных компонент, получились отрицательными. Что в свою очередь в реальном мире невозможно. Цена на товар не может быть отрицательной! Это говорит о том, что модель, построенная на линейной регрессии, не может эффективно работать на этих данных. Для улучшения результатов необходимо использовать ограничения для задания вида аппроксимации, либо перейти к другой параметрической модели.

Если же сравнивать результаты параметрического и непараметрического подходов, то непараметрический подход показывает лучшие результаты, нежели параметрический.

Таким образом, в результате выполнения работы был предложен непараметрический алгоритм прогнозирования стоимости автомобиля, основанный на оценке Надарая-Ватсона, с помощью которого удалось повысить точность решения задачи идентификации в многомерном пространстве в условиях малого объёма данных, содержащих пропуски.

Выводы к главе 3

В результате написания третьей главы можно сделать следующие выводы:

1) на основе эксперимента было произведено сравнение трёх методов заполнения пропусков на выборке, максимально приближенной к реальной;

2) произведено заполнение пропусков методом импутирования линейной регрессией;
3) реализовано сокращение размерности матрицы наблюдений на основе метода главных компонент;
4) построена непараметрическая модель прогнозирования, основанная на оценке Надарая-Ватсона;
5) реализован параметрический метод прогнозирования в студии машинного обучения Azure с помощью линейной регрессии на полученных данных;
6) произведено сравнение предлагаемого метода с существующими.

Был получен непараметрический алгоритм прогнозирования стоимости автомобиля, не имеющий аналогов в мире.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены задачи машинного обучения, с решением которых возможно решение задачи прогнозирования. Восстановление регрессии является лучшим из таких решений. Также исследовалось решение задачи идентификации с помощью параметрического и непараметрического подходов.

Было произведено исследование методов борьбы с пропущенными значениями в матрице наблюдений и реализованы некоторые из них: удаление строк с пропусками, метод Бартлета, Resampling-метод, заполнение на основе линейной регрессионной модели.

Также исследовались методы сокращения размерности матрицы наблюдений и с помощью программного продукта SPSS Statistica был реализован метод главных компонент, который при используемых способах обучения алгоритма показал себя не лучшим образом на представленных данных.

В ходе выполнения работы проанализировались работы по заданной тематике. Рассмотрен существующий способ решения задачи прогнозирования стоимости автомобиля на имеющихся данных с использованием параметрического моделирования в студии машинного обучения Azure.

В работе был предложен непараметрический алгоритм прогнозирования стоимости автомобиля, основанный на оценке Надарая-Ватсона, с помощью которого была увеличена точность решения задачи идентификации в многомерном пространстве наблюдений с пропусками в условиях малого объёма данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) Веников, В.А. Теория подобия и моделирование : учебное пособие / В.А. Веников, Г.В. Веников. – Москва : Высш. шк., 1984. – 243 с.

2) Самарский, А.А. Математическое моделирование. Идеи. Методы. Примеры : монография / А.А. Самарский, А.П. Михайлов. – Москва : Наука. Физмат- лит, 1997. – 320 с.

3) Мухин, О.И. Моделирование систем [Электронный ресурс] / Режим доступа: http://stratum.ac.ru/textbooks/modelir/contents.html4.

4) Орлов, А.И. Теория принятия решений : учебное пособие / А.И. Орлов. - Москва: Издательство «Март», 2004. – 196 с.

5) Олзоева, С.И. Моделирование и расчёт распределённых информационных систем : учебное пособие / С.И. Олзоева. - Улан-Удэ : Издательство ВСГТУ, 2004. - 67 с.

6) Владимирова, Л.П. Прогнозирование и планирование в условиях рынка : учебное пособие / Л.П. Владимирова. – Москва : Издательский Дом «Дашков и Ко», 2001. – 296 с.

7) Слуцкин, Л.Н. Курс МБА по прогнозированию в бизнесе : монография / Л.Н. Слуцкин. – Москва : Альпина Бизнес Букс, 2006. – 38 с.

8) Новикова, Н.В. Прогнозирование национальной экономики : учебно-методическое пособие / Н.В. Новикова, О.Г. Поздеева. – Екатеринбург : Издательство Урал. гос. экон. ун-та, 2007. – 205 с.

9) Эйкхофф, П. Основы идентификации систем управления : монография / П. Эйкхофф. – Москва : Мир, 1975. – 681 с.

10) Заварин, А.Н. Использование априорной информации в непараметрических оценках функции регрессии / А.Н. Заварин. – Москва : Автоматика и телемеханика. – 1985. – №5. – 79-85 с.

11) Цыпкин, Я.З. Информационная теория идентификации / Я.З. Цыпкин. – Москва : Наука. Физматлит, 1995. – 336 с.

12) Катковник, В.Я. Непараметрическая идентификация и сглаживание данных : монография / В.Я. Катковник. – Москва : Наука, 1985. – 336 с.

13) Цыпкин, Я.З. Адаптация и обучение в автоматических системах : монография / Я.З. Цыпкин. – Москва : Наука, 1968. – 400с.

14) Efroimovich, S.Yu. Nonparametric curve estimation. Methods, theory and application : monografh / S.Yu. Efroimovich. – Berlin, New-York: Springer-Verlag, 1999.

15) Шуленин, В.П. Математическая статистика. Ч.1. Параметрическая статистика : учебник / В.П. Шуленин. – Томск: Издательство НТЛ, 2012. – 540 с.

16) Боровков, А.А. Математическая статистика. Оценка параметров. Проверка гипотез : монография / А.А. Боровков. – Москва : Наука, 1984. – 472 с.

17) Медведев, А.В. Непараметрические системы адаптации / А.В. Медведев. – Новосибирск : Наука, 1983. – 173с

18) Шуленин, В.П. Математическая статистика. Ч.2. Непараметрическая статистика: учебник / В.П. Шуленин – Томск: Издательство НТЛ, 2012. – 388 с.

19) Апраушева, Н.Н. Использование непараметрических оценок в регрессионном анализе / Н.Н. Апраушева, В.Д. Конаков – Санкт-Петербург : Заводск. лаб. – 1973. – № 5. – С. 556-569.

20) Медведев, А.В. Адаптация в условиях непараметрической неопределенности / А.В. Медведев // Адаптивные системы и их приложения. – Новосибирск : Наука, 1978. – С. 4-34.

21) Медведев, А.В. Элементы теории непараметрических систем управления / А.В. Медведев // Актуальные проблемы информатики, прикладной математики и механики. Часть 3, Информатика. – Новосибирск-Красноярск: СО РАН, 1996. – С. 87-112.

22) Надарая, Э.А. Непараметрическое оценивание плотности вероятностей и кривой регрессии : учебник / Э.А. Надарая. – Город.: Издательство Тбилисского университета, 1983.

23) Rubin, D.B. Multiple Imputation for Nonresponse in Surveys : manual / D.B. Rubin. - New York : Willey, 1987.

24) Литтл, Р.Дж.А. Статистический анализ данных с пропусками : учебник / Р.А. Литтл, Д.Б. Рубин. - Москва : Наука, 1991. – 198 с.

25) Злоба, Е. Статистические методы восстановления пропущенных данных / Е. Злоба, И. Яцкив. // Computer Modeling & New Technologies.; Vol. 6.2004.

26) Снитюк, В.Е. Эволюционный метод восстановления пропусков в данных [Электронный ресурс] / В.Е. Снитюк. – Режим доступа: http://iissvit.narod.ru/index_a.htm.

27) Снитюк, В.Е. Алгоритм ZetBraid [Электронный ресурс] / В.Е. Снитюк // Информационные интеллектуальные системы. Вып.40, 2008, Режим доступа: http://iissvit.narod.ru/rass/vip40.htm.

28) Rubin, D.B. Multiple imputation after 18+ years. / D.B. Rubin. // Journal of the American Statistical Association, № 91, 1996.

29) Королев, В.Ю. ЕМ – алгоритм, его модификации и их применение к задаче разделелния смесей вероятностных распределений : теоретический обзор / В.Ю. Королев. – Москва : Наука, 2008.

30) Россиев, А.А. Моделирование данных при помощи кривых для восстановления пробелов в данных. В кн. “Методы нейроинформатики” / Под ред. А.Н. Горбаня. - КГТУ: Красноярск, 1998.

31) Тюрин, Ю.Н. Анализ нечисловой информации : учебное пособие / Ю.Н. Тюрин, Б.Г. Литвак, А.И. Орлов, Г.А. Сатаров., Д.С. Шмерлинг. – Москва : Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981.

32) Орлов, А.И. Предельное распределение одной оценки числа базисных функций в регрессии // Прикладной многомерный статистический анализ. Ученые записки по статистике, т.33. – Москва : Наука, 1978.

33) Харман, Г. Современный факторный анализ: учебное пособие / Г. Харман. – Москва : Статистика, 1972.

34) Орлов, А.И. Заметки по теории классификации / А.И. Орлов. – Москва: Социология: методология, методы, математические модели, 1991, № 2.

35) Орлов, А.И. Базовые результаты математической теории классификации / А.И. Орлов. // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 110.

36) Орлов, А.И. Математические методы теории классификации : учебное пособие / А.И. Орлов. – Кубань : Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 95.

37) Терехина, А.Ю. Анализ данных методами многомерного шкалирования : монография / А.Ю. Терехина. – Москва : Наука, 1986. – 205 с.

38) Перекрест, В.Т. Нелинейный типологический анализ социальноэкономической информации: Математические и вычислительные методы / В.Т. Перекрест. – Москва : Наука, 1983.

39) Овечкина, О.О. Агрегация и регрессионный подход к численному моделированию больших данных / О.О. Овечкина. - Научные исследования и разработки молодых учёных, номер 7, 2015.

40) Чернявский, Г.А. Возможности минерагенического прогноза и прогноза сейсмоопасных зон по данным глубинной электроразведки : исследование / И.А.Безрук, В.П. Борисова, Г.А. Чернявский. – Москва : Геофизика. 1995. № 3. С. 26-32.

41) Соколов, О.В. Краткосрочный прогноз притока воды в Бурейское водохранилище на основе модели ECOMAG с использованием метеорологических прогнозов / Ю.Г. Мотовилов, В.В. Балыбердин, Б.И. Гарцман, А.Н.Гельфан, В.М. Морейдо, О.В. Соколов – Москва : Водное хозяйство России: проблемы, технологии, управление. 2017. № 1. С. 78-102.

42) Большаков, А.А. Методы обработки многомерных данных и временных рядов : учебное пособие для студентов вузов, обучающихся по магистерской программе 550209-"Автоматизация науч. исслед., испытаний и эксперимента" направления 550200- "Автоматизация и упр.", по направлениям

230100 (654600)-"Информатика и вычислительная техника" / А. А. Большаков, Р. Н. Каримов.- Москва, 2007.

43) Кошкин, Г. М. Непараметрическая идентификация стохастических объектов : монография / Г. М. Кошкин, И. Г. Пивен. – Хабаравск : Российская акад. наук, Дальневосточное отделение, 2009.

44) Корнеева А.А. Непараметрические модели и алгоритмы управления для многомерных систем с запаздыванием : дис. … канд. техн. наук : 05.13.01 / А.А. Корнеева. – Красноярск, 2014.

45) Родионова, Т.Е. Исследование применимости регрессионного моделирования при решении прецизионных задач астрометрии и небесной механники : дис. … канд. техн. наук : 05.13.18 / Т.Е. Родионова. - Ульяновск, 2003.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра Информатики

УТВЕРЖДАЮ

Заведующий кафедрой

подпись А.С. Кузнецов
инициалы, фамилия

« 21 » июня 2017 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 «Системный анализ и управление»

Непараметрический алгоритм прогнозирования стоимости автомобиля

Руководитель  ст. преподаватель, к.т.н. А. А. Корнеева
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник  И.О.Бессмертный
подпись, дата инициалы, фамилия

Красноярск 2017