МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технологический университет

«СТАНКИН» (ФГБОУ ВО «МГТУ «СТАНКИН»)

Институт
информационных
технологий

Кафедра
информационных технологий
и вычислительных систем

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №1 ПО ДИСЦИПЛИНЕ

«ГИПЕРМЕДИЙНЫЕ СРЕДЫ И ТЕХНОЛОГИИ»

СТУДЕНТА *4* КУРСА *бакалавриата* ГРУППЫ *ИДБ-20-02*
(уровень профессионального образования)

ЕРДОГАНА ДЕНИЗА ЕРДАЛОВИЧА

Направление: 09.03.01 Информатика и вычислительная техника

Профиль подготовки: Программное обеспечение средств вычислительной техники и автоматизированных систем

Отчет сдан «_____» _____2024 г.

Оценка _____

Преподаватель Коган Ю.Г. к.т.н., доцент кафедры ИТиВС _____
(Ф.И.О., должность, степень, звание.) *(подпись)*

МОСКВА 2024

ОГЛАВЛЕНИЕ

ГЛАВА 1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 ИСПОЛЬЗУЕМЫЕ ТЕРМИНЫ

Для грамотного рассмотрения дальнейшего текста, а также для понимания хода повествования, стоит дать определения основных понятий, чтобы читатель полноценно представлял о чём идёт речь.

Нужно отметить, что часть из определений трактуется неоднозначно, указан один из возможных вариантов уместных и подходящих для данной работы.

Если какое-то определение не дано, значит автор считает, что большинство читателей будет иметь понимание о термине.

1.1.1 ОБЩИЕ ТЕРМИНЫ

Data (данные) - совокупность сведений, зафиксированных на определённом носителе в форме, пригодной для постоянного хранения, передачи и обработки.

API - программный интерфейс, который задаёт описание способов взаимодействия одного программного модуля с другим.

Неуспеваемость - это комплексная, суммарная, итоговая неподготовленность студента для дальнейшей учебы.

Information (информация) - данные имеющие ценность для субъекта.

Big Data (Большие данные) - структурированные или неструктурированные массивы данных большого объема.

1.1.2 МАТЕМАТИЧЕСКИЕ ТЕРМИНЫ

Генеральная совокупность – совокупность всех возможных объектов, которые подлежат изучению в пределах объекта исследования.

Выборочная совокупность (выборка) - множество объектов определённым способом выбранных из генеральной совокупности для использования в исследовании.

1.1.3 MACHINE LEARNING ЧАСТЬ

Machine Learning (машинное обучение) - наука создания алгоритмов (моделей), которые улучшаются благодаря опыту.

1.1.4 DEEP LEARNING ЧАСТЬ

Deep Learning (глубокое обучение) - наука создания алгоритмов (моделей), которые используют данные как опыт.

1.2 МОТИВАЦИЯ АВТОРА

Выбор темы данной ВКР не случаен. Автор два с половиной года изучает машинное обучение и предметные области науки Data Sciencея. Были изучены материалы разных типов связанные с ВКР: курсы (на различных языках), книги, видеоматериал. Имеется опыт построения моделей на известной платформе Kaggle. Что касается математического аппарата, то его автор развивает по сей день при помощи академической учёбы (школа, университет) и увлечением сопуствующим материалом в свободное время. По поводу уровня владения современными инструментами проблем не имеется, автор обучается на IT-направлении, что подразумевает под собой программирование и изучение электронно-вычислительных устройств, также

автор дополнительного проходит образовательную программу в школе программирования (соответствующие работы в приложении).

Данная ВКР будет использоваться создателем в дальнейшем не только в рамках высшего образования, но и в роли пет-проекта для профессиональной деятельности. Так что качество обеспечивается как личностными интересами, так карьерными.

1.3 ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ П. О.

Данное API можно будет использовать как открытый веб-сервис для пользователей, которые хотят узнать информацию о себе и своей успеваемости. Для этого пользователю нужно будет ввести необходимую информацию и последовать инструкции для получения статистических данных.

Если сервис будет востребован у пользователей, то будет осуществлена его модернизация и дальнейшая модификация для использования его как полноценного продукта на рынке.

Что касается масштабируемости, то имеются прилагающиеся возможности. Например, реализовать портативную версию для мобильных устройств.

Продукт имеет практически безграничный потенциал к развитию, так как современные тенденции связанные с машинным обучением развиваются и будут развиваться.

1.4 АНАЛИЗ ЛИТЕРАТУРЫ

Для снижения влияния субъективной точки зрения автора в исследование рассматривается разного рода литература, причём чем

разнообразнее она будет относительно подхода рассмотрения проблемы, тем лучше удастся раскрыть проблему.

Однако имеется риск зашумление данных избыточной информацией или усиления влияния определённых признаков за счёт транзитивной зависимости дополнительных признаков, так нужно выверить идеальную грань в этом вопросе.

Для отслеживания общих ключевых моментов будем обращаться к уже имеющимся исследованиям, чтобы иметь какой-то baseline. Рассматриваться результаты той или иной работы не будут, задача смотреть на то, что актуального может пригодиться в ВКР.

Нужно отметить сразу, что работа происходит дело с несбалансированной информацией в связи с ограниченностью способов проведения опроса. Так очевидно что часть вузов имеют больше приоритет и популярность, в связи с этим студенты престижных вузов будут иметь преимущество в этом плане, именно поэтому автор старался собираться информацию с вузов разных регионов и вузов с разным уровнем.

Рассматриваемые источники:

- Источник № [6]: выявляется “теснота связи среднего балла успеваемости студентов и суммарного балла ЕГЭ для каждого курса свыше 38%”. Нужно отметить, что, возможно, хотелось большой связи, однако надо учитывать, что исследование было сделано на большой выборки, разных направлений, курсов. Из-за этого прогностическая сила падает, однако отрицать выведенную в исследовании связь нецелесообразно.

  В связи с этим связь результатов ЕГЭ и успеваемости студентов позволяет сделать вывод о валидности ЕГЭ как показателя успеваемости выпускников.

- Источник № [7]: была выявлена интересная связь по итогам набора 2009 года. Снижении качественных показателей бюджетных студентов, при

увеличении их количества. Автор раскрывает данную свзязь тем, что чем больше конкуренция на место, тем более трудолюбивые и талантливые попадают на бюджетные места (Прицнип сильнейшего).

- Источник № [8]: автором обобщаются данные научных работ, посвященных факторам и предикторам академической успешности. Формулируется вывод о смещении в психологопедагогических исследованиях значимости когнитивных факторов успеваемости в сторону личностных и социальных факторов. Исследование проводились в ФГБОУ ВО «Липецкий государственный педагогический университет имени П.П. Семенова-Тян-Шанского» города Липецк. В работе были выявлены следующие признаки:

  - мотивация;

  - пол;

  - тревожность;

  - настойчивость;

  - обустройство в социуме (данная фича раскрывается так: не для кого не секрет, что студенты помогают друг другу как работами, так объяснением материала).

Также были выявлены внешние влияющие факторы:

- мнение родственников о необходимости высшего образования (преимущественно родители);

- наличие высшего образования у родителей;

- какой доход у Вашей семье;

Однако сразу отметим, что наиважнейшими являются личностные качества, но исключать внешние не стоит, так как они, возможно, будут влиять на успеваемость студента.

ГЛАВА 2. КОНСТРУИРОВАНИЕ ПРОЦЕССА

2.1 СТРУКТУРА ПРОЕКТА

Данную работу удобно рассматривать как Data Science проект. Все ключевые этапы практически идентичны в обоих случаях. При таком подходе будет иметься чёткий план действий в детерминированном порядке.

В Data Science-проекте имеется три основных раздела, в каждом из которых 3 этапа. Рассмотрим основные разделы:

- Работа с требованиями: на этом этапе необходимо вникнуть в постановку задачи, понять, какой результат требуется получить от проекта, узнать про участников. В соответствии с определенной задачей нужно решить, какой метод использовать для решения задачи. Результатом этого шага будут требования к данным: что может понадобиться для успешного решения;

- Работа с данными: необходимо приступить к поиску данных для решения задачи: узнать, какие источники доступны, и сформировать выборку, с которой в дальнейшем будет осуществляться работа. После того как данные собраны, необходимо провести ряд исследований, чтобы лучше понимать, как устроена выборка:

  - Исследовать: центральное положение, вариабельность;
  - Выявить корреляции между признаками;
  - Построить графики распределения.

После этого этапа можно приступать к подготовке данных. Как правило, этот этап самый трудоемкий процесс. В зависимости от того, насколько качественно он выполнен, зависит успех всего проекта;

- Разработка и внедрение: после того, как данные готовы, можно приступать к разработке и внедрению. Программируем модель, прогоняем на обучающей выборке, проверяем на тестовой, если результат устраивает, то демонстрируем

заказчику, внедряем, собираем фидбэк, если нет, то дорабатываем до удовлетворительного результата.

Теперь рассмотрим подробнее продемонстрированные выше разделы по этапам (См. таблица № 1):

Таблица 1. Описание основных этапов Data Science проекта

Название этапа	Описание этапа
Понимание задачи	Необходимо определить цель исследования: что является проблемой? Почему проблема должна быть решена? Кого затрагивает проблема? Главное: по каким метрикам будет оцениваться успешность проекта? Иными словами, необходимо выявить цель.
Аналитический подход	Нужно выбрать аналитический подход для решения бизнес-задачи. Выбор подхода зависит от того, какой тип ответа нужно получить в итоге: - если ответ должен быть вида да/нет, то нужен классификатор; - если ответ значение численного признака, то используются регрессионные модели; - промежуточный вариант перечисленных деревья решений; - если нужно определить вероятность, необходимо использовать предиктивную модель; - если необходимо выявить связи, используется дескриптивный подход.

Требования к данным	Когда определена цель исследования и выбран подход, необходимо определиться с тем, какие данные позволят дать искомый ответ. Нужно подготовить требования к данным: контент, форматы, источники.
Сбор данных	На этом этапе выполняется сбор данных из имеющихся источников: убеждаемся, что источники доступны, надежны и могут быть использованы для получения искомых данных в требуемом качестве. После необходимо понять, получили ли данные, какие хотели. На этой стадии можно пересмотреть требования к данным и принять решения о необходимости дополнительных данных. Могут быть выявлены лакуны в данных и составлен план, как их закрыть или найти замену.
Анализ данных	Анализ данных включает в себя все работы по конструированию выборки. На этом этапе необходимо получить ответ на вопрос: репрезентативны ли собранные данные относительно поставленной задачи? Здесь используется описательная статистика. Она применяется ко всем переменным, которые будут использоваться в выбранной модели: - исследуется центральное положение; - ищутся выбросы и выполняется оценка вариабельности; - строятся гистограммы распределения переменных; - визуализируются данные; - выполняется попарное сравнение: вычисляются корреляции между переменными, чтобы определить, какие из них связаны и насколько. Если найдутся значительные

	корреляции между переменными, какие то из них могут быть отброшены, как избыточные.
Подготовка данных	На этом этапе перерабатываем данные в такую форму, чтобы с ними было удобно работать: удаляем дубликаты, обрабатываем отсутствующие или неверные данные, проверяем и при необходимости исправляем ошибки форматирования. Также на этом этапе конструируем набор факторов, с которым на следующих этапах будет работать машинное обучение: проводим извлечение и отбор признаков, которые потенциально помогут решить бизнес-задачу. Ошибки на этом этапе могут оказаться критическими для всего проекта, поэтому к нему стоит отнестись особенно внимательно: избыточное количество признаков может привести к тому, что модель будет переобучена, а недостаточное — к тому, что модель будет недообучена.
Построение модели	Выбор модели, как можно было заметить, осуществляется в самом начале работы и зависит от бизнес-задачи. Таким образом, когда тип модели определен и имеется обучающая выборка, аналитик разрабатывает модель и проверяет, как она работает на созданном на этапе 6 наборе признаков.
Применение модели	Применение модели идет в тесной связке с собственно построением модели: вычисления чередуются с настройкой модели. На этом этапе мы должны ответить на вопрос, отвечает ли построенная модель бизнес-задаче. Вычисление модели имеет две фазы: проводятся диагностические измерения, которые помогают понять,

	работает ли модель, так как задумано. Если используется предиктивная модель, может использоваться дерево решений, чтобы понять, что выдача модели соответствует изначальному плану. На второй фазе проводится проверка статистической значимости гипотезы. Она необходима, чтобы убедиться, что данные в модели правильно используются и интерпретируются и полученный результат выходит за пределы статистической погрешности.
Внедрение	Если модель дает нам удовлетворительный ответ на поставленный вопрос, этот ответ должен начать приносить пользу. Когда модель разработана, и аналитик уверен в результате своей работы, необходимо познакомить заказчика с разработанным инструментом. Имеет смысл привлечь не только владельца продукта, но и других заинтересованных лиц: маркетинг, разработчиков, системные администраторы: всех, кто хоть как то может оказать влияние на дальнейшее использование результатов проекта. Далее необходимо переходить к внедрению. Внедрение может происходить поэтапно, например, на ограниченную группу пользователей или в тестовом окружении. Также необходимо наладить систему фидбэка, чтобы отслеживать, насколько успешно разработанная модель справляется с поставленной задачей. Через некоторое время этот фидбэк будет полезен для того, чтобы усовершенствовать модель.

Даже внедренная модель никогда не может считаться идеальной. Если этапы визуализировать, то получится следующий рисунок (См. рисунок № 1):

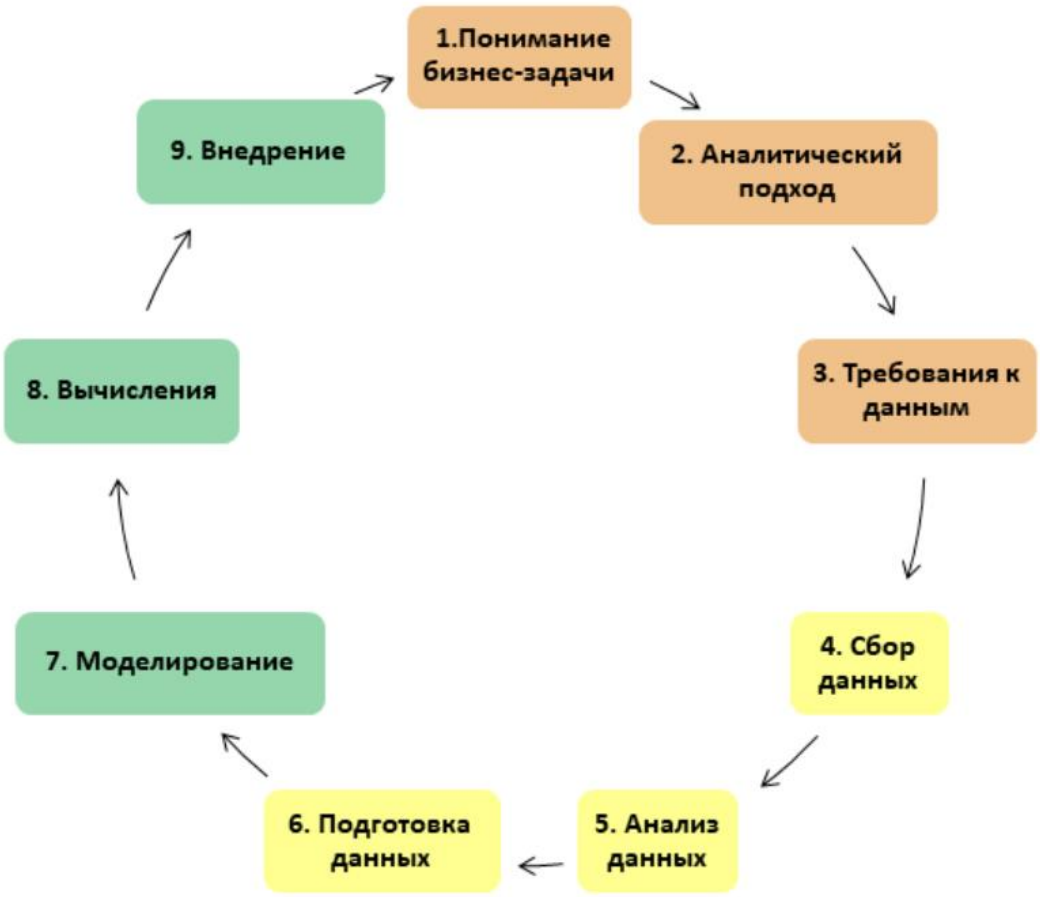

Рис. 1 Основные этапы Data Science проекта

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Толстова Ю. Н. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ СОЦИОЛОГОВ : учеб. пособие / Толстова Ю. Н. — 2-е изд.. — Москва: Юрайт, 2021 — 258 с.;
2. Вьюгин В. В, МАТЕМАТИЧЕСКИЕ ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ И ПРОГНОЗИРОВАНИЯ [Текст] / Вьюгин В. В, — 1-ое изд.. — Москва: , 2013-2020 — 484 с.;
3. Савельев В. Статистика и котики [Текст] / Савельев В. — 1-ое изд.. — Москва: АСТ, 2018 — 170 с.;
4. Академия наук СССР Математические методы в социологическом исследовании [Текст] / Академия наук СССР — 1-ое изд.. — Москва: Наука, 1981 — 33 с.;
5. Бурков А. МАШИННОЕ ОБУЧЕНИЕ без лишних слов [Текст] / Бурков А. — 2-ое изд.. — Санкт-Петербург: Питер, 2020 — 192 с.
6. Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Социальные науки, 2017, № 1 (45), с. 171–177 URL: http://www.unn.ru/pages/e-library/vestnik_soc/18115942_2017_-_1(45)_unicode/23.pdf Дата публикации: 02.02.2017. Режим доступа: гость (дата обращения: 24.02.2024).
7. Жданов, Д. Н. АНАЛИЗ УСПЕВАЕМОСТИ СТУДЕНТОВ ДЛЯ ОЦЕНКИ ДЕЯТЕЛЬНОСТИ КУРАТОРА ГОУ ВПО «Алтайский государственный технический университет им. И. И. Ползунова» г. Барнаул.
   URL: http://elib.altstu.ru/disser/conferenc/2010/01/pdf/317zhdanov.pdf.
   Режим доступа: гостевой (дата обращения: 07.03.2024).
8. Дормидонтов, Р. А. Проблема успеваемости и успешности обучающихся в свете социальных изменений развития общества и образовательных систем / Р. А. Дормидонтов // Мир науки. Педагогика и психология. —

2022. — Т. 10. — № 5. — URL: https://mir-nauki.com/PDF/27PDMN522.pdf. Режим доступа: гостевой (дата обращения: 01.03.2024).