Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Факультет «Высшая школа экономики и управления»
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент, начальник УИ
ФГАОУ ВО «ЮУрГУ (НИУ)»
_____ (В.Г. Раенко)
«____» _____ 2019 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н.,
с.н.с.
_____ (Б.М. Суховилов)
«____» _____ 2019 г.

Разработка математических моделей для оценки времени испытаний автомобильных сборок на примере компании Мерседес

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2019.301.ПЗ ВКР

Руководитель проекта, д.т.н.
_____ (В.В. Мокеев)
«____» _____ 2019 г.

Автор проекта,
студент группы ЭУ– 244
_____ (А.В. Зайцев)
«____» _____ 2019 г.

Нормоконтролер, доцент
_____ (Е.В. Бунова)
«____» _____ 2019 г.

Челябинск 2019

АННОТАЦИЯ

Зайцев А.В. Разработка математических моделей для оценки времени испытаний автомобильных сборок на примере компании Мерседес. Челябинск: ЮУрГУ, ЭУ-244, 2019. – 67 с., 15 ил., 8 табл., библиогр. список – 13 наим.

Выпускная квалификационная работа посвящена разработке математических моделей для оценки времени испытаний автомобильных сборок на примере компании Мерседес.

В работе представлены материалы исследования процесса испытания автомобильных сборок. Проведен анализ машинного обучения, а также обоснован выбор, почему именно его следует использовать для прогнозирования времени тестирования автомобильных сборок. Представлены методы машинного обучения, а также сделан выбор в сторону метода, использованного при прогнозировании сбоев. В работе присутствует описание выбранных методов, исходные данные предоставленные компанией Мерседес, предварительная обработка данных, а также результаты проведенной работы. Описана дорожная карта коммерциализации проекта, создан сайт по предоставлению услуги прогнозирования времени тестирования автомобильных сборок. Создана модель машинного обучения целью, которой является предсказание времени тестирования автомобиля на основе его конфигурации, с заданной точностью. Рассчитан медиаплан и ценовая политика коммерциализации проекта.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы

Безопасность и надежность - это решающий шаг в процессе производства автомобилей. Каждый новый дизайн автомобиля должен пройти тщательную оценку, прежде чем он войдет на потребительский рынок. Тестирование может занять много времени и потребовать больших затрат, так как полная проверка систем автомобиля требует подвергнуть автомобиль воздействию всех ситуаций, с которыми он столкнется в своем намеренном использовании. Прогнозирование общего времени прохождения испытания на испытания затруднено, потому что для каждой модели требуется другая конфигурация тестового стенда. Mercedes-Benz является первопроходцем многочисленных функций безопасности и технических характеристик транспортных средств и предлагает ряд индивидуальных вариантов для каждой модели. Каждая возможная комбинация транспортных средств должна проходить те же тщательные испытания, чтобы гарантировать, что автомобиль достаточно прочен, чтобы обеспечить безопасность пассажиров и выдерживать ежедневные нагрузки. Большой набор опций, предлагаемых Mercedes, - это большое количество тестов для инженеров компании. Увеличение количества тестов приводит к увеличению времени, затрачиваемого на стенд, увеличению затрат компании и производству углекислого газа, загрязняющего окружающую среду. Таким образом, компания довольно сильно заинтересована в решении данной проблемы, и работа является вполне актуальной на сегодняшний день.

В выпускной квалификационной работе описывается подход к решению проблемы испытания автомобильных сборок компании Мерседес. Минимизация издержек процесса без ущерба для качества очень важна для компании. На тестовом стенде автомобильные сборки проходят этап контроля качеста, очень важно на данном этапе выявить все недочеты, для гарантии безопастности и качества конечной продукции. Оптимальное время испытаний должно является уникальным для каждой автомобильной сборки. В случае отклонения от оптимального времени испытания в сторону уменьшения, появляется риск понижения точности испы-

таний, и как следствие снижения качества конечного продукта, в обратном же случае отклонения от оптимального времени испытания в сторону увеличения, увеличиваются издержки компании без повышения качества испытаний, что является недопустимым, а так же бесполезным. Исходя из этого, появилась потребность, в создании IT решения для оценки времени испытаний автомобильных сборок.

Мы представляем наши выводы из набора данных. Исследуем проблемы, с которыми сталкиваются из-за размера набора данных, типа записанных данных и алгоритмов машинного обучения, которые подходят для такого рода задач. В разделе I описан процесс испытания автомобильных сборок, факторы влияющие на время испытаний, в разделе II рассмотрены регрессионные методы используемые в машинном обучении, включая алгоритм линейная регрессия, случайный лес, экстра-деревьев, градиентный бустинг, а также метрика качества и процедура – кросс-валидация. В разделе III представлен набор данных и его предварительная обработка, исследование эффективности методов: линейная регрессия, случайный лес, экстра-деревьев, градиентный бустинг. В разделе IV будет рассмотрена коммерциализация проекта

Актуальность темы обусловлена необходимостью прогнозирования времени испытаний автомобильных сборок на примере компании Мерседес. Благодаря предоставленным данным, может быть построена более разумная система оценки времени испытаний, и сформирован более оптимальный план испытательных работ. Все это позволит снизить издержки в данном сегменте и увеличить прибыль

Основной целью работы является – Снижение затрат компании в области тестирования готовой продукции(автомобилей) без понижения качества испытаний.

Чтобы достичь поставленную цель, необходимо решить следующие задачи:

– проанализировать процесс испытания автомобильных сборок

– проанализировать виды испытаний, и зависимость длительности испытаний от всевозможных факторов

– проанализировать использование методов машинного обучения для решения регрессионных задач

– объяснить выбор использованных метрик качества и процедуры кросс-валидации;

– проанализировать предоставленный набор данных;

– провести предварительную обработку данных;

– исследовать эффективность методов машинного обучения для решения данной регрессионной задачи.

– спрогнозировать время испытаний автомобильных сборок на примере компании Мерседес

– разработать коммерциализацию проекта

Научной новизной является использование метода градиентного бустинга дляпрогнозирования времени испытаний автомобильных сборок

Практическая значимость – использование данного подхода позволяет снизить эксплуатационные расходы и увеличить прибыль производственных предприятий.

1 ПРОЦЕСС ИСПЫТАНИЯ АВТОМОБИЛЬНЫХ СБОРОК

1.1 Виды испытаний

Испытание автомобилей отличают по назначению, способам проведения и по объектам испытания. Проводятся испытания, как макетных образцов, так и опытных. Опытные и макетные образцы и их модификаций подвергаются доводке, заблаговременным и приемочным проверкам. Машины текущего производства проходят контрольные, ресурсные, приемочные и сертификационные тестирования, кроме того тесты на надежность. Образчики всех без исключения транспортных средств на каждой стадии их разработки и изготовления могут пройти эксплуатационные, экспериментальные и специализированные, детерминирующие, проверки. По методам, условиям и пункту проведения тестирования разделяют на стендовые, полигонные с применением различных разновидностей дорог, водоемов, ванн, подъемов, неровностей и т. д., дорожные с регламентацией качества дорог общего применения, эксплуатационные и тестирования в разных природных условиях.

1.2 Длительность испытаний

Длительность испытаний АТС определяется отрезком времени от начала доставки выделенных объектов на место проведения до момента, когда полученная информация становится достаточной для выполнения целей и задач, намеченных рабочей программой. Основные факторы, влияющие на длительность испытаний:

− объём и организация подготовки объектов испытаний, средств и оборудования;

− план проведения испытаний;

− эффективность методов экспериментальных работ;

− режимы внешних воздействий;

− обоснованность и эффективность показателей, критериев и свойств, определяемых при испытаниях;

− качество испытательного оборудования, материально-технического обеспечения, квалификация персонала;

− наличие и эффективное использование дополнительной информации об испытываемых объектах.

В зависимости от этих факторов наибольшую длительность имеют эксплуатационные испытания в условиях рядового использования автомобилей у потребителей. Сравнительно с этим видом испытаний и рассматривается их ускорение. Ближайшими по содержанию, объёму и качеству информации к испытаниям в рядовой эксплуатации, но существенно ускоренными, являются полигонные испытания. Еще большее сокращение длительности достигается при стендовых испытаниях, но преимущественно по отдельным элементам конструкции или характеристикам. Эффективность этих испытаний проявляется на этапе доводки конструкции. Основным фактором ускорения стендовых испытаний является непрерывность процессов нагружения и иных внешних повреждающих воздействий. Оборудование современных полигонов предусматривает проведение и стендовых, и лабораторно-дорожных, и ходовых испытаний с выполнением всех рабочих (технологических) функций. Для ускорения полигонных испытаний используется влияние всех перечисленных выше факторов на длительность получения необходимой информации. При лабораторно-дорожных испытаниях на полигоне сокращение длительности достигается уплотнением подготовительных и организационных работ, стабильностью технологии и технической базы, повышением производительности труда за счёт поточных методов их проведения. Наиболее длительной и трудоёмкой частью полигонных испытаний являются пробеговые (с выполнением рабочих технологических функций). Сокращение времени и трудовых затрат на их проведение является определяющим направлением эффективности испытаний полнокомплектных автотранспортных средств. Реализация этого направления (повышение темпов испытательных работ) зависит от оборудования полигона и квалификации испытателей. Основным фактором ускорения испытаний на полигоне является усиление режимов внешних воздействий, формируемых

в испытательных пробегах на специально обустроенных дорогах. Здесь следует подчеркнуть особо, что при испытаниях автомобильной техники решающую роль играют испытательные пробеги. Эта часть испытаний даёт интегральные оценки испытываемой модели − надёжности, эксплуатационной технологичности, фактического проявления всех заложенных в конструкцию потребительских свойств, включая оценку гарантийных сроков, сроков службы до списания. Только в реальных пробегах окончательно оценивается пригодность данной машины для выполнения заданных функций. Ускоренные полигонные испытания без форсирования нагружения получили название нормальных или рядовых. Нормальные пробеговые испытания проводятся в полигонных условиях на основе традиционных порядков их осуществления, сложившихся до создания автополигонов, хотя организация их на новой базе сама по себе давала повышение темпов, сокращение сроков, более высокую стабильность условий в сравнении с пробеговыми испытаниями на сети дорог общего пользования в различных условиях. В основу планирования нормальных пробеговых полигонных испытаний кладется исследование условий работы АТС по назначению, оценка и анализ режимов нагружения агрегатов, узлов и деталей в эксплуатации прототипов, аналогов. Так как подавляющее большинство типов АТС имеют широкий диапазон случайных режимов нагружения и обстоятельств движения, то для воспроизведения их на автополигоне подбираются типизированные условия эксплуатации. Например, для автомобилей общетранспортного назначения выделяют режимы городских, магистральных и горных условий перевозки. Для воспроизведения таким образом типизированных условий на полигоне подбирается комплекс дорог, соответствующих по характерным признакам (ровности, сопротивлению качения, распределения подъемов и спусков) типизированным условиям эксплуатации. Устанавливается регламентированный пробег с нормативными значениями средних скоростей движения на каждой испытательной дороге, распределением по ним общего пробега в долях, чередованием движения и остановок, продолжительностью перерывов (по санитарногигиеническим нормам труда испытателей). Отработка таких нор-

мативов полигонных испытаний создает предпосылки для 36 их ускорения с помощью форсирования воздействия внешних факторов, в первую очередь, специальных испытательных дорог интенсивного и направленного нагружения, специальных испытательных каналов, камер, а также организации пробеговых испытаний в характерных климатических зонах страны. Пробег в предварительных испытаниях планируется в два этапа. Первый назначается в объеме гарантируемой заводом-изготовителем протяженности, второй - до исчерпания ресурса (до капитального ремонта). Причем, как правило, второй этап разбивается на части, равные гарантийному пробегу. В каждой части пробега соблюдаются условия первого этапа.

1.3 Постановка задачи

Процесс испытания автомобильных сборок является процессом выходного контроля качества конечной продукции. Сложно недооценить важность данного процесса. На этом этапе автомобили проходят все возможные тестирования гарантирующие водителю безопасность, конструкция автомобильных сборок проходит эксплуатационные испытания гарантирующие долговечность и надежность конечного продукта. Оптимальное время тестирование автомобильных сборок является уникальным для каждой из них. Отклонение от оптимального времени испытаний в меньшую сторону приведет к снижению точности, или качества испытаний, что приведет к выпуску менее качественной продукции, а так же возможно и менее безопасной, что не допустимо, потому что это может подвергнуть опасности жизни потребителя. Превышение же оптимального времени испытаний приводит к повышению издержек компании в данном процессе, без повышения качества испытаний, что является упущенной экономической выгодой. Поэтому перед компанией Мерседес встала проблема оценки оптимального времени тестирования автомобильных сборок. Компанией Мерседес был подготовлен набор данных разделенный на обучающую и тестовую выборку несущий в себе конфи-

гурации автомобильных сборок и в обучающей выборке целевую переменную y несущую в себе время испытаний. Для создания модели машинного обучения которая будет прогнозировать оптимальное время тестирования автомобильных сборок.

1.4 Выводы по главе 1

Мы рассмотрели процесс испытания автомобильных сборок. Это производственный процесс, в котором производится контроль качества выпускаемой продукции.

Были рассмотрены факторы влияющие на время испытаний автомобильных сборок и методы снижения данного времени. Рассмотрены наиболее важные правила, которые необходимо учитывать при создании системы контроля качества.

Исходя из исследования предметной области, мы пришли к выводу о том, что для решения задачи прогнозирования времени испытания автомобильных сборок эффективно использовать методы машинного обучения.

2 ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ РЕГРЕССИОННЫХ ЗАДАЧ

2.1 Регрессионные методы используемые в машинном обучении

Задачи в машинном обучении, разделяются на широкие категории. Данные категории базируются на том, как обучение получено либо как обратная взаимосвязь по обучению предоставляется разработанной системе.

Двумя наиболее популярными методами машинного обучения считаются контролируемое обучение, которое обучает методы на базе примерных входных и выходных данных, помеченных людьми, и неконтролируемое обучение, которое обеспечивает алгоритм без участия помеченных данных, для того чтобы предоставить возможность ему находить структуру в собственных входных сведений. Давай проанализируем данные способы наиболее детально.

Кто-то в силах задать вопрос: «По какой причине машины обязаны обучаться? Почему бы не проектировать машины таким образом, чтобы они трудились так, как хотелось бы?» Имеется ряд факторов, согласно которым машинное обучение немаловажно. Безусловно, мы уже затрагивали, то что достижение обучения на машинах может помочь нам осознать, как животные и человечество обучаются. Однако имеется и значимые технические предпосылки.

Вот некоторые из них:

1. Некоторые задачи никак не могут быть отчетливо определены, кроме как на примере; в таком случае мы имели возможность бы указать пары ввода / вывода, однако никак не точную взаимосвязь между входами и желаемыми выходами. Мы хотели бы, чтобы машины имели возможность адаптировать собственную внутреннюю структуру с целью извлечения верных выходных данных с целью значительного количества выборочных входов и, таким образом, ограничивать их функцию ввода / вывода с целью аппроксимации отношения, неявного в примерах.

2. Возможно, что из числа огромных куч данных спрятаны значимые взаимосвязи и корреляции. Методы машинного обучения зачастую имеют все шансы применяться с целью извлечения данных взаимоотношений (интеллектуальный анализ данных).

дизайнеры зачастую производят машины, которые никак не функционируют так, как хотелось бы, в условиях, в которых они используются. По сути, определенные свойства рабочей среды могут быть не полностью известны во время исследования. Методы машинного обучения могут быть применены с целью усовершенствования имеющихся конструкций машин.

объем доступных знаний об определенных задачах может быть слишком большим для явного кодирования людьми. Машины, которые изучают это знание постепенно, могут захватить его больше, чем люди захотят записать;

среда меняется со временем. Машины, которые могут адаптироваться к изменяющейся среде, уменьшат необходимость в постоянном проектировании;

новые знания о задачах постоянно открываются людьми.

Изменения словарного запаса. В мире постоянно происходит поток новых событий. Продолжать модернизацию систем искусственного интеллекта для соответствия новым знаниям нецелесообразно, но методы машинного обучения могут быть в состоянии отследить большую часть этого.

Обучение, как и интеллект, охватывает такой широкий спектр процессов, которые трудно точно определить. Словарное определение включает в себя такие фразы, как «чтобы получить знания, или понимание, или навыки, путем изучения, обучения или опыта» и «изменение поведенческой тенденции с помощью опыта». Зоологи и психологи изучают обучение на животных и людях. В этой книге мы сосредоточимся на обучении в машинах. Существует несколько параллелей между обучением животных и машин. Конечно, многие методы машинного обучения основаны на попытках психологов уточнить свои теории обучения животных и человека с помощью вычислительных моделей. Представляется также вероятным,

что концепции и методы, изучаемые исследователями в области машинного обучения, могут пролить свет на некоторые аспекты биологического обучения.

Что касается машин, мы можем очень широко сказать, что машина учится всякий раз, когда она меняет свою структуру, программу или данные (на основе своих входных данных или в ответ на внешнюю информацию) таким образом, что улучшается ожидаемая в будущем производительность. Некоторые из этих изменений, такие как добавление записи в базу данных, удобно укладываются в область других дисциплин и не обязательно лучше понимаются, поскольку их называют обучением. Но, например, когда производительность машины для распознавания речи улучшается после прослушивания нескольких образцов речи человека, в этом случае мы чувствуем себя вполне оправданными, чтобы сказать, что машина научилась. Машинное обучение обычно относится к изменениям в системах, которые выполняют задачи, связанные с искусственным интеллектом (ИИ). Такие задачи включают распознавание, диагностику, планирование, управление роботом, прогнозирование и т.д.

«Изменения» могут быть либо улучшением уже работающих систем, либо синтезом новых систем. Эта система воспринимает и моделирует свое окружение и вычисляет соответствующие действия, возможно, предвидя их последствия. Изменения, внесенные в любой из компонентов, могут учитываться как обучение. Различные механизмы обучения могут быть использованы в зависимости от того, какая подсистема изменяется. Мы будем изучать несколько различных методов обучения в этой работе. Различают два метода обучения контролируемое и обучение без учителя, рассмотрим каждое поподробнее.

2.1.1 Линейная регрессия

В статистике линейная регрессия-это линейный подход к моделированию связи между скалярным откликом (или зависимой переменной) и одной или несколькими независимыми переменными (или независимыми переменными). Случай одной независимой переменной называется простой линейной регрессией. Для

нескольких независимых переменных процесс называется множественной линейной регрессией. Этот термин отличается от многомерной линейной регрессии, где предсказываются несколько коррелированных зависимых переменных, а не одна скалярная переменная.

В линейной регрессии отношения моделируются с использованием линейных предикторных функций, неизвестные параметры модели которых оцениваются по данным. Такие модели называются линейными моделями. Чаще всего условное среднее ответа, заданное значениями независимых переменных (или предикторов), считается аффинной функцией этих значений; реже используется условная медиана или какой-либо другой квантиль. Как и все формы регрессионного анализа, линейная регрессия фокусируется на условном распределении вероятностей ответа с учетом значений предикторов, а не на совместном распределении вероятностей всех этих переменных, которое является областью многомерного анализа.

Линейная регрессия была первым типом регрессионного анализа, который изучаясь, широко использовался в практических приложениях. это связано с тем, что модели, линейно зависящие от неизвестных параметров, легче приспосабливать, чем модели, нелинейно связанные с их параметрами, и потому, что статистические свойства полученных оценок легче определить.

Линейная регрессия имеет много практических применений. Большинство приложений относятся к одной из следующих двух широких категорий:

Если целью является прогнозирование, или прогнозирование, или уменьшение ошибок, линейная регрессия может использоваться для подгонки прогнозной модели к наблюдаемому набору данных значений отклика и независимых переменных. После разработки такой модели, если дополнительные значения независимых переменных собираются без сопутствующего значения отклика, подобранная модель может быть использована для прогнозирования отклика.

Если цель состоит в объяснении вариации переменной ответа, которая может быть отнесена к вариации независимых переменных, линейный регрессионный анализ может применяться для количественной оценки силы связи между ответом и независимыми переменными и, в частности, для определения того, могут ли некоторые независимые переменные вообще не иметь линейной связи с ответом, или для определения того, какие подмножества независимых переменных могут содержать избыточную информацию об ответе.

Модели линейной регрессии часто приспосабливаются, используя подход наименьших квадратов, но они могут также быть приспособлены другими способами, такими как минимизация "отсутствия подгонки" в некоторой другой норме (как с регрессией наименьших абсолютных отклонений), или минимизируя штрафную версию функции стоимости наименьших квадратов как в регрессии гребня (штраф L2-норма) и лассо (штраф L1-норма). И наоборот, подход наименьших квадратов можно использовать для подгонки моделей, которые не являются линейными моделями. Таким образом, хотя термины "наименьшие квадраты" и "линейная модель" тесно связаны, они не являются синонимами.

2.1.2 Случайный лес (Random forest)

Случайный лес — один из самых потрясающих алгоритмов машинного обучения, придуманные Лео Брейманом и Адель Катлер ещё в прошлом веке. Он дошёл до нас в «первозданном виде» (никакие эвристики не смогли его существенно улучшить) и является одним из немногих универсальных алгоритмов. Универсальность заключается, во-первых, в том, что он хорош во многих задачах (по оценкам, 70% из встречающихся на практике, если не учитывать задачи с изображениями), во-вторых, в том, что есть случайные леса для решения задач классификации, регрессии, кластеризации, поиска аномалий, селекции признаков и т.д.

RF (random forest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).

Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления — свои случайные признаки).

Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Понятно, что такая схема построения соответствует главному принципу ансамблирования (построению алгоритма машинного обучения на базе нескольких, в данном случае решающих деревьев): базовые алгоритмы должны быть хорошими и разнообразными (поэтому каждое дерево строится на своей обучающей выборке и при выборе расщеплений есть элемент случайности).

Метод RF хорош ещё тем, что при построении леса параллельно может вычисляться т.н. oob-оценка качества алгоритма (которая очень точная и получается не в ущерб разделения на обучение/тест), oob-ответы алгоритмы (ответы, которые выдавал бы алгоритм на обучающей выборке, если бы «обучался не на ней»), оцениваются важности признаков. Также не стоит забывать про полный перебор значений параметров (если объектов в задаче не очень много).

2.1.3 Экстра-деревья (Extra-Tree)

Метод Extra-Tree (стоящий для extremely randomized trees) был предложен с главной целью дальнейшего построения дерева рандомизации в контексте число-

вых признаков ввода, где выбор оптимальной точки пересечения отвечает для значительной части дисперсии индуцированного дерева.

Что касается случайных лесов, метод исключает идею использования загрузочных копий учебного образца, и вместо того, чтобы пытаться найти оптимальную точку пересечения для каждой из случайно выбранных признаков K на каждом узле, он выбирает точку пересечения наугад.

Эта идея довольно продуктивна в контексте многих проблем, характеризующихся большим числом числовых признаков, изменяющихся более или менее непрерывно: она часто приводит к повышенной точности благодаря ее сглаживанию и в то же время значительно снижает вычислительное бремя, связанное с определением оптимальных срезы в стандартных деревьях и в случайных лесах.

С статистической точки зрения, отбрасывание идеи начальной загрузки приводит к преимуществу с точки зрения смещения, тогда как рандомизация с режущей средой часто является отличным эффектом уменьшения дисперсии. Этот метод позволил получить самые современные результаты в нескольких многомерных сложных задачах.

С функциональной точки зрения, метод Extra-Tree создает кусочно-полилинейные аппроксимации, а не кусочно-постоянные из случайных лесов.

2.1.4 Градиентный бустинг

XGBoost - это контролируемый алгоритм обучения, который реализует процесс, называемый boosting, чтобы дать точные модели. Boosting относится к методу обучения ансамблю для построения многих моделей последовательно, причем каждая новая модель пытается исправить недостатки предыдущей модели. В повышении дерева каждая новая модель, добавленная в ансамбль, является деревом решений. XGBoost обеспечивает параллельное наращивание дерева (также известное как GBDT, GBM), которое быстро и точно решает многие проблемы с

наукой о данных. Для многих проблем XGBoost - одна из лучших рамок ускорителя градиента (GBM) сегодня.

Возможности XGBoost – особенности модели и системные функции

Реализация модели поддерживает особенности реализации scikit-learn и R с новыми дополнениями, такими как регуляризация. Поддерживаются три основные формы повышения градиента: Алгоритм Gradient Boosting также называется градиентной машиной повышения, включая скорость обучения; Stochastic Gradient Boosting с суб-выборкой в строке, столбце и столбце на каждый уровень разделения; Регулярное усиление градиента с регуляцией L1 и L2.Библиотека предоставляет систему для использования в различных вычислительных средах, не в последнюю очередь: Параллелизация построения дерева с использованием всех ваших ядер процессора во время обучения; Распределенные вычисления для обучения очень крупных моделей с использованием кластера машин; Вне корпоративного вычисления для очень больших наборов данных, которые не вписываются в память; Кэш Оптимизация структуры данных и алгоритма для наилучшего использования аппаратного обеспечения.

Реализация алгоритма была разработана для эффективности вычислительных ресурсов времени и памяти. Цель проекта заключалась в том, чтобы наилучшим образом использовать имеющиеся ресурсы для обучения модели. Некоторые ключевые функции реализации алгоритма включают: Редкая реализация Aware с автоматической обработкой отсутствующих значений данных; Блочная структура для поддержки распараллеливания конструкции дерева;

Продолжение обучения, чтобы вы могли еще больше повысить уже установленную модель для новых данных.

После того как мы изучили алгоритмы машинного обучения стоит сделать выбор в пользу того алгоритма, который будет наилучшим выбором для решения нашей задачи. Несомненно, для начала следует разобраться в тех данных и задачах, которые перед нами поставлены

2.2 Выводы по главе 2

Методы машинного обучения постоянно совершенствуются. Мы рассмотрели несколько популярных алгоритмов, которые используются в машинном обучении:

- Линейная регрессия
- Случайный лес – RF (random forest);
- Extra-Trees (стоящий для extremely randomized trees);
- XGBoost;

Использован самый распространенный метод для оценки производительности алгоритма машинного обучения – это использование различных наборов данных train/test, а также деление на k фолды. А также, рассмотрена кросс-валидация, которую мы будем использовать в нашей работе. Это перекрестная проверка – подход, который можно использовать для оценки производительности алгоритма машинного обучения с меньшей дисперсией, чем в случае разделения набора из одного train.

3 ИССЛЕДОВАНИЕ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ВРЕМЕНИ ИСПЫТАНИЙ АВТОМОБИЛЬНЫХ СБОРОК КОМПАНИИ МЕРСЕДЕС

3.1 Подготовка данных для подачи в модель машинного обучения

Для решения данной проблемы Mercedes-Benz предоставляют два файла данных. Учебный набор данных и набор данных тестирования. Оба файла представлены в формате с разделителями-запятыми (CSV) и доступны для загрузки на странице данных конкурса Kaggle. Данные по подготовке и тестированию содержат 4209 тестов на автомобиль, полученных Mercedes для различных конфигураций транспортных средств. Данные обучения также содержат целевую переменную или продолжительность тестирования в секундах для каждого теста транспортного средства. Для тестовых данных не предусмотрена цель, так как длительность тестирования известна только Mercedes и используется для определения лучшей модели. Каждое испытание транспортного средства определяется конфигурацией транспортного средства, которая кодируется набором функций. Оба набора для обучения и тестирования содержат 376 различных характеристик автомобиля с такими именами, как «X0», «X1», «X2» и т. Д. Все функции были анонимизированы, что означает, что они не имеют физического представления. Описание данных указывает на то, что характеристиками автомобиля являются параметры конфигурации, такие как настройка подвески, адаптивный круиз-контроль, полный привод и ряд различных вариантов, которые вместе определяют модель автомобиля. Существует 8 категориальных функций со значениями, закодированными как строки, такие как «a», «b», «c» и т. Д. Другие 368 функций являются бинарными, то есть они либо имеют значение 1, либо 0. Каждое испытание транспортного средства также был присвоен идентификатор, который не рассматривался как функция для этого анализа. Ниже изображен образ репрезентативного подмножества учебных данных:

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	...	X375	X376	X377	X378	X379	X380
0	0	130.81	k	v	at	a	d	u	j	o	...	0	0	1	0	0	0
1	6	88.53	k	t	av	e	d	y	l	o	...	1	0	0	0	0	0
2	7	76.26	az	w	n	c	d	x	j	x	...	0	0	0	0	0	0
3	9	80.62	az	t	n	f	d	x	l	e	...	0	0	0	0	0	0
4	13	78.02	az	v	n	f	d	h	d	n	...	0	0	0	0	0	0
5	18	92.93	t	b	e	c	d	g	h	s	...	0	0	1	0	0	0
6	24	128.76	al	r	e	f	d	f	h	s	...	0	0	0	0	0	0

Рисунок 1: Пример данных обучения

Хотя данные были очищены Mercedes до того, как они были доступны для участия в соревнованиях, и поэтому нет ошибок или недостающих записей, данные могут по-прежнему содержать выбросы в отношении времени тестирования транспортного средства. Эти выбросы могут быть достоверными данными, но достаточно экстремальными, чтобы повлиять на производительность модели.

Для наглядности был построен график (рисунок 2) показывающие с какого диапазона значений в данных идут выбросы

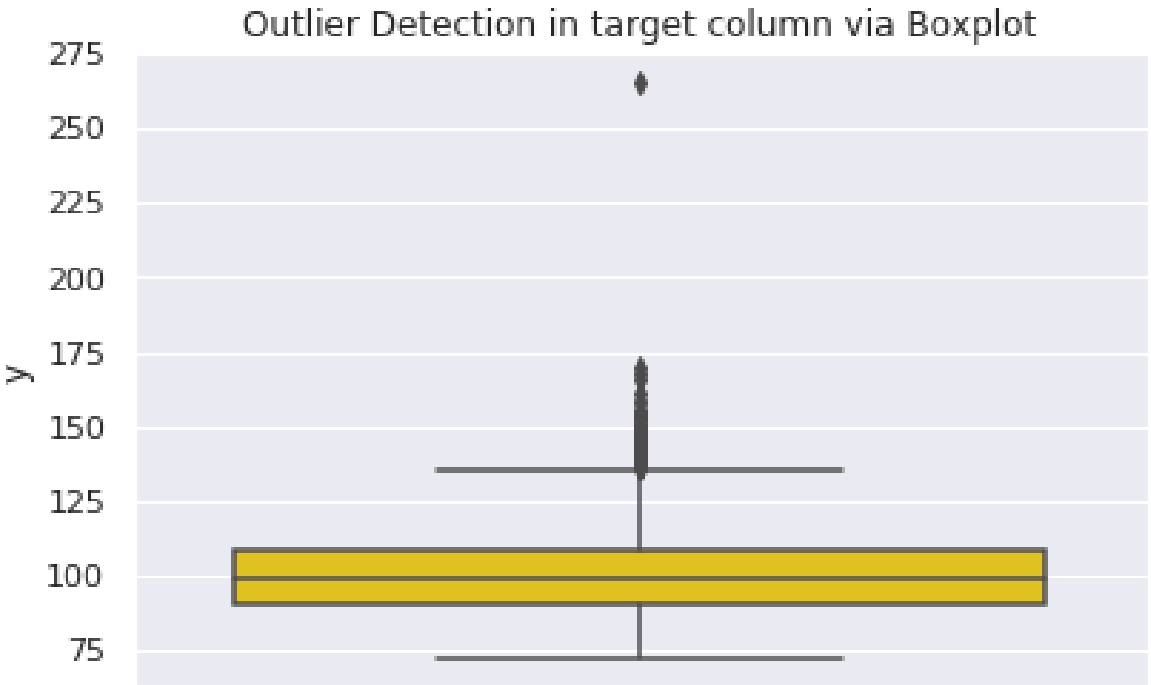

Рисунок 2: Поиск выбросов

Исходя из рисунка 2 наглядно видно, что выбросы превышают значение прибл. 137.5

Таким образом, было произведено удаление выбросов на основе вышеуказанной информации и установка 137.5 в качестве порогового значения.

Чтобы модель машинного обучения обрабатывала категориальные переменные, они должны быть закодированы в горячем состоянии. Это означает, что уникальные категории, содержащиеся в категориальной переменной, преобразуются в набор новых переменных. Каждому экземпляру присваивается значение для новой переменной, соответствующей ее исходной категориальной переменной, и нулю во всех других новых переменных. Это лучше всего иллюстрируется рисунком 3.

	A	B	C	D	E	F	G	H	I
1	Original data:			One-hot encoding format:					
2	id	Color		id	White	Red	Black	Purple	Gold
3	1	White		1	1	0	0	0	0
4	2	Red		2	0	1	0	0	0
5	3	Black		3	0	0	1	0	0
6	4	Purple		4	0	0	0	1	0
7	5	Gold		5	0	0	0	0	1
8									

Рисунок 3: Однократное кодирование

Однократное кодирование преобразует все функции в двоичные значения. После того, как данные тестирования и обучения были разогреты, два набора данных были выровнены, чтобы исключить любые функции, присутствующие в одном наборе данных, но не в другом. Это необходимо, потому что модель не будет знать, как реагировать на нее, если столкнулась с особенностью набора те-

стов, который он не видел в данных обучения. После однократного кодирования и выравнивания данных обучения и тестирования в обоих наборах данных было в общей сложности 553 двоичных элемента.

Первым аспектом данных для исследования была целевая переменная, продолжительность тестирования транспортного средства. В приведенном ниже графике показаны все время тестирования, расположенное от кратчайшего до самого длинного слева направо. Четыре отклонения можно увидеть справа с наивысшим значением, показанным красным. Построение данных демонстрирует более экстремальный характер этой точки данных, чем анализ чисел.

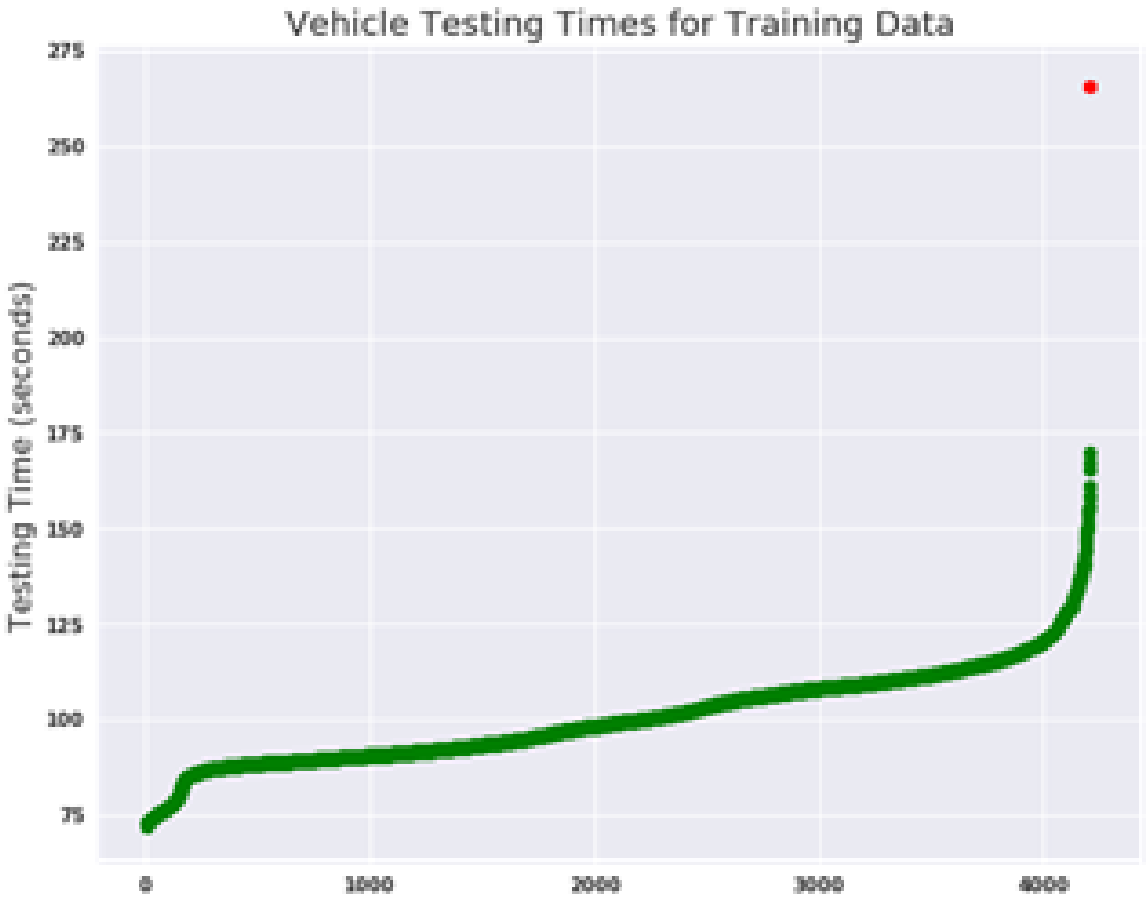

Рисунок 4: Время тестирования транспортных средств для данных обучения

Таким образом удаленная точка данных была удалена для обучения, чтобы лучше представлять большую часть данных.Если посмотреть далее на числовые

данные для набора тестов, было определено 12 двоичных переменных, где значения были либо 1, либо 0. Поскольку в этих функциях нет изменений, они не имеют прогностической способности, и поэтому эти 12 функций могут быть удалены.

В идеале, построенная модель должна иметь небольшие и смещение и дисперсию поэтому был произведен анализ данных критериев

Удаление столбцов с нулевой дисперсией показана на рисунке 5

```
temp = []
for i in df_num.columns:
    if df[i].var()==0:
        temp.append(i)
print(len(temp))
print(temp)
```

```
13
['X11', 'X93', 'X107', 'X233', 'X235', 'X268', 'X289', 'X290', 'X29
3', 'X297', 'X330', 'X339', 'X347']
```

Рисунок 5: Удаление столбцов с нулевой дисперсией

Установка порога 0,01 для дисперсии для каждого столбца и последующее удаление не вошедшего в порог. Удаленные столбцы также удаляются из всех временных кадров данных.

```
count=0
low_var_col=[]
for i in test.columns:
    if test[i].dtype == 'int64':
        if test[i].var()<0.01:
            low_var_col.append(i)
            count+=1
print(count)

df.drop(low_var_col,axis=1,inplace=True)
df_num.drop(low_var_col,axis=1,inplace=True)
test.drop(low_var_col,axis=1,inplace=True)
```

```
146
```

Рисунок 6: Отбор столбцов для удаления

Таким образом, есть 146 столбцов для удаления.

Исследование данных подтвердило необходимость уменьшения размерности. К счастью, одним из распространенных методов сокращения количества функций, анализа основных компонентов (PCA), является неконтролируемый метод, который не требует понимания физического представления функций.

Выводы из исследования данных заключаются в следующем:

Необходимо удалить один выброс, определяемый продолжительностью тестирования транспортного средства.

12 двоичных переменных не кодируют никакой информации и должны быть удалены

Данные с нулевой дисперсией так же удаляются

·Обоснованно неконтролируемое уменьшение размерности (PCA)

3.2 Метрика качества

Оценочной метрикой для соревнования является мера R^2, известная как коэффициент детерминации. R^2 является мерой качества модели, которая используется для предсказания одной непрерывной переменной из ряда других переменных. В нем описывается величина вариации зависимой переменной, в этом случае время тестирования транспортного средства в секундах на основе независимых переменных, в данном случае комбинация пользовательских функций транспортного средства, которые могут быть объяснены моделью. Его часто интерпретируют как процент изменения целей, который объясняется функциями. Таким образом, значение $R^2$0,6 указывает на то, что 60% изменения времени тестирования может быть объяснено изменением в настройке транспортного средства. Остальные 40% дисперсии либо не учитываются моделью, или из-за скрытых переменных, которые не были включены в данные. Коэффициент детерминации выражается математически в уравнении 1.

$$R^2 = \left(\frac{n*(\sum x*y)-(\sum x)(\sum y)}{n*[(\sum x^2)-(\sum x)^2]*[n*(\sum y^2)-(\sum y)^2]}\right)^2 \quad [1]$$

где n - количество экземпляров (тесты транспортных средств),

x - предсказание для экземпляра (предсказанное время тестирования),

y - известное значение истины для экземпляра (известное время тестирования в секундах).

R^2 of 0 может быть достигнуто простым рисованием прямой линии по данным при среднем значении целевой переменной. Наилучший возможный коэффициент детерминации - 1.0, который указывает, что модель объясняет всю дисперсию переменной ответа в терминах входных переменных.

Коэффициент детерминации является подходящей метрикой для проблемы, поскольку цель, определенная Mercedes-Benz, заключается в создании модели, ко-

торая может определять время тестирования транспортного средства. Mercedes интересуется тем, почему разные транспортные средства принимают разные времена для тестирования и как это можно представить в модели машинного обучения. Поэтому алгоритм, который наилучшим образом объясняет вариацию времени тестирования, будет оптимальной моделью машинного обучения для задачи. Коэффициент детерминации является общей метрикой, используемой в регрессионных задачах, и реализуется в Scikit-Learn, где это оценочная оценка по умолчанию для регрессора. Коэффициент детерминации для данных обучения может быть установлен на этапе оценки модели, поскольку данные обучения включают в себя известные целевые значения; однако R^2для данных тестирования может быть найдено только путем представления прогнозов от модели к конкуренции.

3.4 Исследование эффективности построенной модели

Окончательная модель, созданная для этого проекта, сочетает в себе множество различных методов машинного обучения. На самом высоком уровне конечная модель представляет собой взвешенное голосование между двумя промежуточными моделями. Первая промежуточная модель представляет собой ансамблевый метод, известный как усиление градиента [15], который работает, создавая множество простых регрессоров друг на друга, чтобы создать окончательный сильный регресс. Вторая промежуточная модель представляет собой сложную модель, в которой еще один метод ансамбля, дополнительный лес деревьев, строится на вершине регуляризованной линейной регрессии. Архитектура показана ниже на рисунке 7.

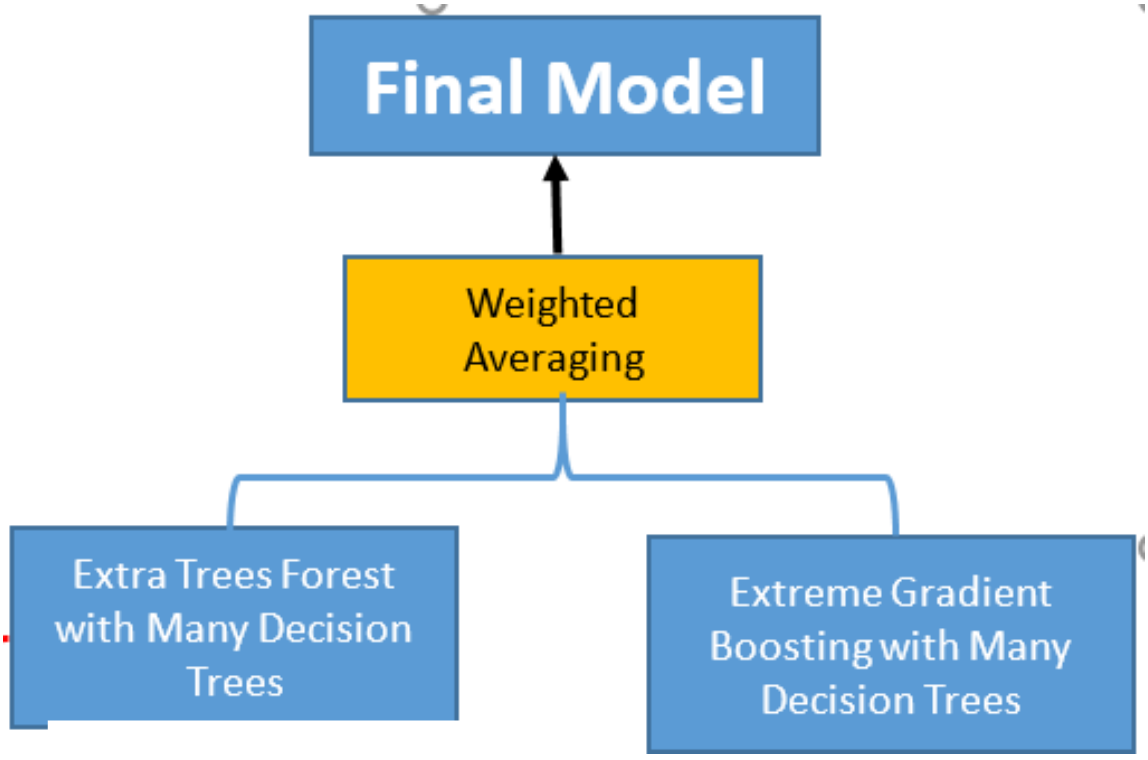

Рисунок 7: Архитектура полной модели

Окончательная модель была получена на основе исследований из книг, дискуссионного форума Kaggle и документов, в которых обсуждались преимущества и недостатки различных моделей. Почти все топ-исполнительные модели в соревновании Kaggle использовали метод XGBoost, усредненный или уложенный по-

верх других методов ансамбля. Именно здесь я получил архитектуру верхнего уровня. С этого момента речь шла о том, какие модели лучше всего дополняли друг друга. На основе эталонной модели я увидел, что линейная регрессия выполнена достаточно хорошо, но, как правило, перегружает данные обучения. Поэтому моя сложная модель включала линейную регрессию с регуляризацией, чтобы уменьшить количество переобучения. Выбор леса с дополнительными деревьями был сделан путем определения эффективности методов ансамбля, уложенных поверх Линейной регрессии. Основываясь на индивидуальной производительности каждой промежуточной модели, было очевидно, что объединение двух прогнозов в средневзвешенном показателе поможет повысить надежность модели. Полный процесс разработки алгоритмов и их оптимизация для проблемы обсуждается вРаздел реализации .

Чтобы объяснить, как работает модель, лучше всего начинать с самого низкого уровня - линейную регрессию с эластичной сетчатой регуляризацией. Полные работы линейной регрессии объясняются в Benchmark Modelсечение и эластичная сетка являются одним из многих способов регуляризации линейной модели. Регуляризация можно рассматривать как ограничение сложности модели, чтобы предотвратить переобучение данных обучения. Overfitting означает, что модель «запоминает» данные обучения, приводящие к плохому обобщению в новых экземплярах, которые ранее не видели. Говорят, что модель, которая перегружает данные обучения, имеет высокую дисперсию и низкую предвзятость. Регуляризация может уменьшить дисперсию модели за счет уменьшения степеней свободы внутри модели и может быть выполнена на линейной модели за счет уменьшения величины параметров модели. Регуляция эластичной сети применяет штраф к каждому параметру модели в функции стоимости, что побуждает модель выбирать более мелкие параметры во время обучения.

Идея сложной модели относительно проста: результаты (предсказания) первой модели используются в качестве входных данных во вторую модель. В этом случае во время обучения предсказания (длительность тестирования транспорт-

ных средств) из линейной регрессии с регуляризацией подаются в качестве входных данных в рекреационный лес с дополнительными деревьями наряду с известными целевыми метками. Таким образом, регрессор Extra Trees изучает прогнозы, основанные на предсказании предыдущей модели и известных истинных значений. Регрессор Extra Trees - это так называемый ансамблевый метод. Он работает, объединяя несколько более простых моделей в одну сложную модель, тем самым уменьшая дисперсию от одной модели и создавая лучшие прогнозы. Регрессия Extra Trees построена из множества алгоритмов дерева решений. Дерево решений создает блок-схему (дерево) вопросов (как правило, в виде пороговых значений для функций) во время обучения, чтобы разбить точки данных на все меньшие бункеры, каждый с другим прогнозируемым целевым значением. Во время тестирования Дерево решений перемещается по блок-схеме по одному узлу за раз и помещает точку данных в соответствующий бит на основе особенностей экземпляра. Ниже представлена простая модель дерева решений.

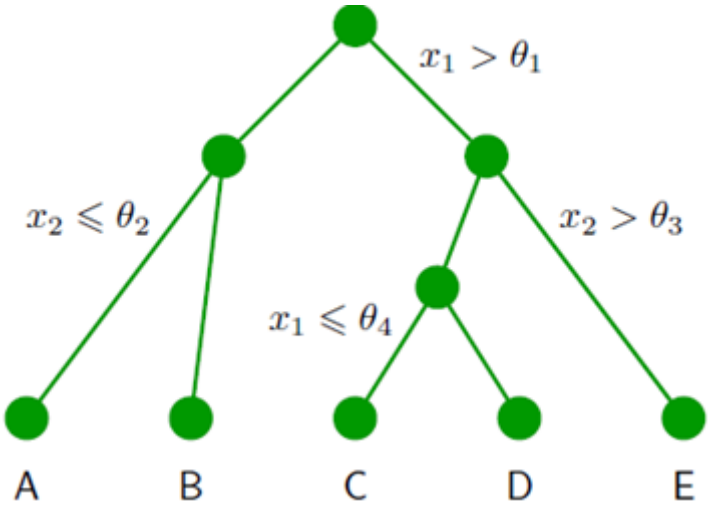

Рисунок 8: Модель дерева решений [18]

В этом случае x-члены будут функциями, θ-терминами будут пороговые значения, установленные во время обучения, и буквы будут конечной предсказанной целью. Регрессион Extra Trees обучает много деревьев решений и принимает во внимание все деревья, чтобы определить окончательное предсказание для каждого экземпляра.

Сложенная модель образует половину конечной модели. Другая половина состоит из другого метода ансамбля, называемого Extreme Gradient Boosting. Принцип усиления экстремального градиента такой же, как и для регрессионного режима Extra Trees, за исключением того, что в этом случае отдельные деревья принятия решений обучаются последовательно, и каждое дерево обучается остаткам предыдущего дерева или разнице между предсказанием дерева и истинным значением ,

Обучаясь остаткам или ошибкам в предсказаниях, каждое последующее дерево становится лучше прогнозировать самые сложные для прогнозирования экземпляры (с наибольшими остатками) и со временем, ансамбль становится более мощным, чем один классификатор. Окончательное предсказание представляет собой средневзвешенное значение по всем отдельным прогнозам, причем более уверенные деревья получают более высокий вес. Повышение градиента стало одним из алгоритмов выбора для соревнований по компьютерному обучению из-за его прогностической способности. Деревья принятия решений составляют основу обоих ансамблевых методов, потому что они относительно быстро тренируются и имеют хорошо установленные параметры по умолчанию.

На верхнем уровне конечная модель принимает среднее значение прогнозов от каждой промежуточной модели. Утяжеление, данное каждой модели, можно определить с помощью итерационного процесса настройки взвешивания и определения производительности модели. Во время обучения предварительно обработанные данные будут переданы обеим промежуточным моделям. Модель Extreme Gradient изучит пороги для каждого листа в лесу деревьев решений, которые он обучает. Сложенная модель сначала передает данные обучения через линейную регрессию, где модель будет изучать параметры (взвешивание), применяемые к каждой функции, тогда линейная регрессия сделает прогноз для каждой точки тренировки и передаст это на вход в Дополнительные деревья деревьев вместе с известными целевыми значениями. Регрессор Extra Trees также сформирует собственный лес деревьев решений с порогами для каждого раскола, определенного

во время обучения. При тестировании каждый новый экземпляр будет передан обеим промежуточным моделям. В случае сложенной модели предсказание будет производиться с помощью линейной регрессии, а затем регрессивный режим Extra Trees сделает прогноз на основе результата линейной регрессии. Модель Gradient Boosting будет генерировать собственное предсказание. Тогда общее предсказание будет среднее из двух промежуточных моделей. Архитектура модели относительно сложна, но жестокий (но дружелюбный) конкурс на Kaggle поощряет разработку уникальных моделей для достижения небольшого повышения производительности. каждый новый экземпляр будет передан обеим промежуточным моделям. В случае сложенной модели предсказание будет производиться с помощью линейной регрессии, а затем регрессивный режим Extra Trees сделает прогноз на основе результата линейной регрессии. Модель Gradient Boosting будет генерировать собственное предсказание. Тогда общее предсказание будет среднее из двух промежуточных моделей. Архитектура модели относительно сложна, но жестокий (но дружелюбный) конкурс на Kaggle поощряет разработку уникальных моделей для достижения небольшого повышения производительности. каждый новый экземпляр будет передан обеим промежуточным моделям. В случае сложенной модели предсказание будет производиться с помощью линейной регрессии, а затем регрессивный режим Extra Trees сделает прогноз на основе результата линейной регрессии. Модель Gradient Boosting будет генерировать собственное предсказание. Тогда общее предсказание будет среднее из двух промежуточных моделей. Архитектура модели относительно сложна, но жестокий (но дружелюбный) конкурс на Kaggle поощряет разработку уникальных моделей для достижения небольшого повышения производительности. Модель Gradient Boosting будет генерировать собственное предсказание. Тогда общее предсказание будет среднее из двух промежуточных моделей. Архитектура модели относительно сложна, но жестокий (но дружелюбный) конкурс на Kaggle поощряет разработку уникальных моделей для достижения небольшого повышения производи-

тельности. Модель Gradient Boosting будет генерировать собственное предсказание. Тогда общее предсказание будет среднее из двух промежуточных моделей.

3.5 Оптимизация гиперпараметров модели

3.5.1 Оптимизация параметров метода градиентного бустинга

Методом перебора параметров было выявлено что на точность модели градиентного бустинга влияют следующие параметры:

«ETA»- Размер шага, используемый в обновлении, предотвращает переоснащение. После каждого шага повышения мы можем напрямую получать веса новых функций и etaсокращать веса функций, чтобы сделать процесс повышения более консервативным.

«Subsample»- Соотношение подвыборок учебных экземпляров. Установка его на 0,5 означает, что XGBoost будет случайным образом отбирать половину обучающих данных до выращивания деревьев. и это предотвратит переоснащение. Подсэмплинг будет происходить один раз в каждой итерации повышения.

«MaxDepth»- Максимальная глубина дерева. Увеличение этого значения сделает модель более сложной и более подходящей. 0 принимается только в lossguidedрастущей политике, когда tree_method установлен как, histи это указывает на отсутствие ограничений по глубине. Помните, что XGBoost активно потребляет память при обучении глубокому дереву.

На рисунке 9 показан подбор параметра «ETA», Диапазон данного параметра может быть задан от 0 до единицы, но предварительный анализ выявил что искать оптимальное значение нужно в диапазоне от 0.001 до 0.009.

Как видно из графика максимальная точность по r2score(valid)(точности модели на валидационной выборке по метрике r2score skikitlearn) достигается при значении «ETA» 0.003, дальнейшее увеличение параметра приводит к устойчивому снижению точности модели.

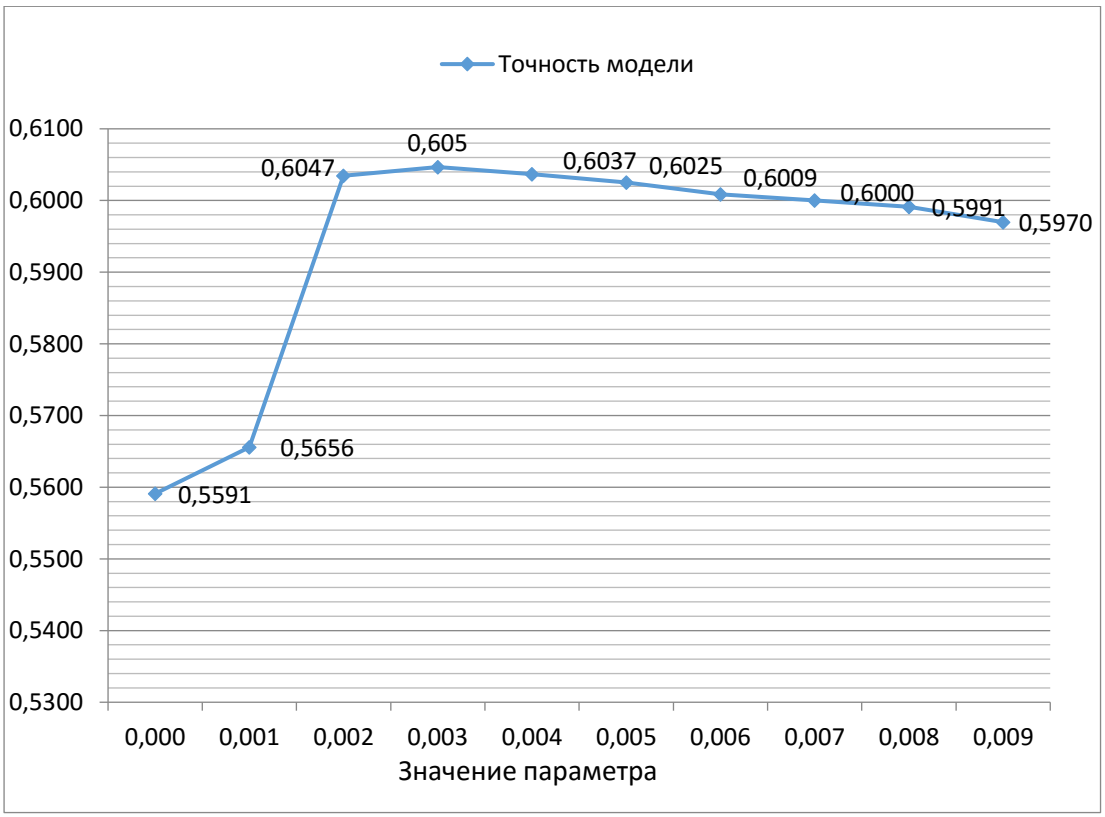

Рисунок 9 – Оптимизация параметра “ETA”

По такому же принципу было найдено оптимальное значение subsample, диапазон данного гиперпараметра может быть от 0 до 1, оптиальное значения для максимальной точности модели является значение 0.7.

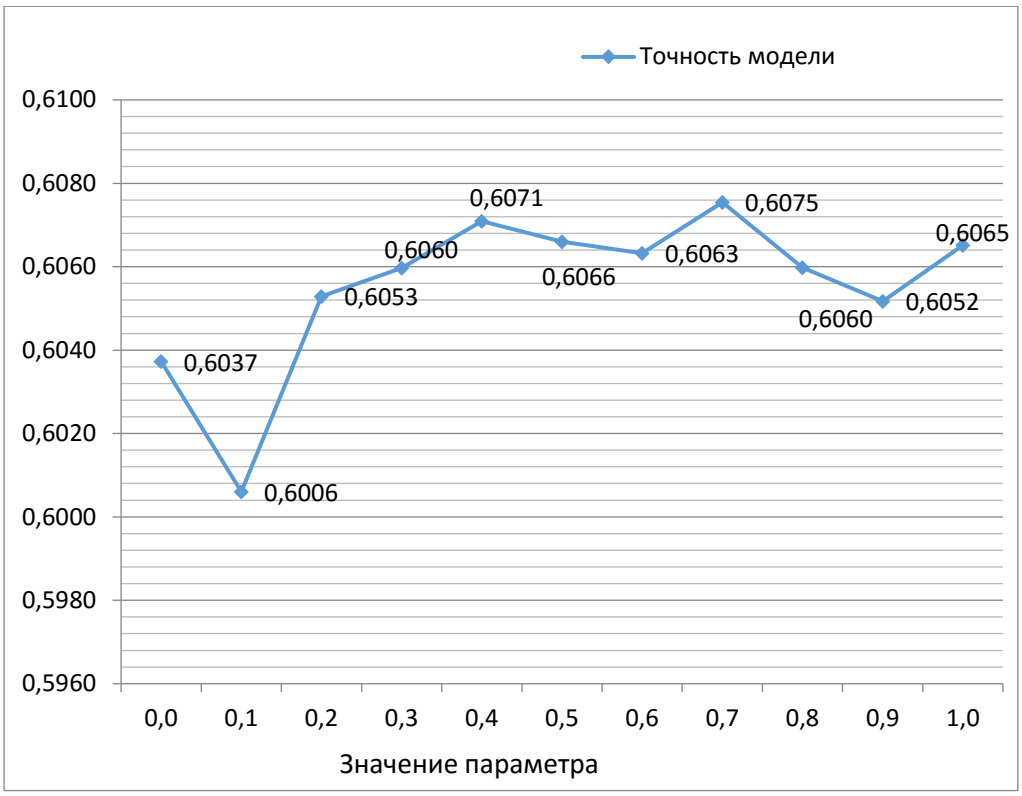

Рисунок 10 – Оптимизация параметра «sublample»

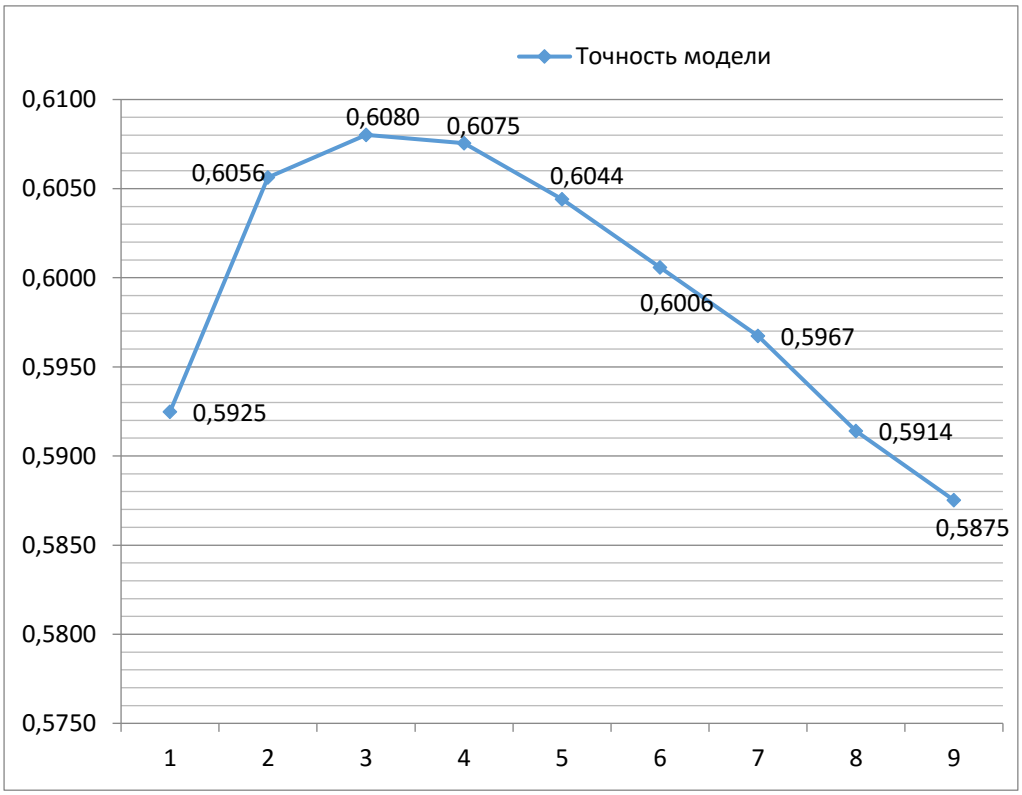

Рисунок 11 – Оптимизация параметра «MaxDepth»

На рисунке 11 выявлено оптимальное значение параметра «MaxDepth» и это значение 3, данный параметр можно увеличивать до бесконечности но мы видим устойчивое снижение точности после значения 3.

Остальные параметры не отраженные в графиках были либо установлены в значение по умолчанию в Scikit-learn, либо их изменение не оказывало влияния на результативность модели.

3.5.2 Оптимизация параметров модели Экстра деревьев (Extra-Trees regresion)

Для оптимизации параметров Экстра деревьев (Extra-Tree) был использован функционал GridSearchCV.

CV в GridSearchCV означает перекрестное подтверждение, которое является методом предотвращения переобучения данных обучения. Концепция перекрестной проверки заключается в том, что данные сначала разделяются на набор для обучения и тестирования. Затем данные обучения разбиваются на «n» меньшие подмножества, известные как складки. Каждая итерация через проверку креста, алгоритм обучается на n-1 на этих подмножествах и оценивается на n-м подмножестве. Конечная оценка алгоритма - это средний балл по всем складкам. Это предотвращает переобучение, благодаря чему модель очень хорошо изучает данные обучения, но затем не может обобщить на новые данные, которые они не видели раньше.

После завершения итераций кросс-валидации модель оценивается в последний раз на тестовом наборе для определения производительности на ранее невидимом наборе данных. Кросс-валидация сочетается с Grid Search для оптимизации алгоритма для данной проблемы, а также для того, чтобы дисперсия модели была не слишком высокой, чтобы модель могла обобщаться на новые экземпляры. GridSearchCV позволяет оценивать широкий диапазон моделей более эффективно, чем вручную проверять каждый вариант.

Окончательные гиперпараметры для модели экстра деревьев, представлены в таблице 3. Любые гиперпараметры, не указанные в таблице, были установлены в значение по умолчанию в Scikit-learn.

Таблица 3: Гиперпараметры модели

Параметр	Экстремальные деревья (Extra Trees regression)
Количество деревьев	500
Максимальное количество функций	0.6
начальная загрузка	false
Минимальное разделение образцов	20

GridSearchCV выполнялся на сложной модели регрессии Extra Trees, уложенной на регуляризованной линейной регрессии) индивидуально.

. Основными параметрами, скорректированными для промежуточной модели, были отношение l1 и допуски для линейной регрессии и максимальные характеристики, минимальное количество выборок на лист и количество оценок для регрессии Extra Trees. Отношение l1 для Elastic Net контролирует штраф, назначенный параметрам модели (он указывает смесь между Ridge и Lasso Regression), а допустимость - минимальный размер шага для продолжения работы алгоритма. В терминах регрессионного режима Extra Trees максимальными характеристиками являются количество функций, которые каждое дерево учитывает, минимальное количество выборок на лист - это минимальное количество точек данных, которое должно быть в каждом листовом узле, а число оценок - это число деревьев решений в лесу. L1_ratio был увеличен для Elastic Net, что приводит к увеличению штрафа (если отношение равно 1,0, то Elastic Net эквивалентно регрессии Lasso, которая стремится устранить наименее важные функции, устанавливая весы, близкие к нулю). Было уменьшено максимальное количество функций для

Regresor Extra Trees, что означает, что модели не нужно было использовать все функции. Обе эти корректировки подсказывают мне, что 140 основных компонентов, возможно, было слишком много, потому что обе модели выполняли неявный выбор признаков через гиперпараметры. Тем не менее, сохранение слишком большого количества основных компонентов, а затем наличие некоторых, не используемых моделью, предпочтительнее не иметь достаточных возможностей и, следовательно, отбрасывать полезную информацию. то Elastic Net эквивалентна регрессии Lasso, которая имеет тенденцию устранять наименее важные функции, устанавливая весовые коэффициенты, близкие к нулю). Было уменьшено максимальное количество функций для Regresor Extra Trees, что означает, что модели не нужно было использовать все функции. Обе эти корректировки подсказывают мне, что 140 основных компонентов, возможно, было слишком много, потому что обе модели выполняли неявный выбор признаков через гиперпараметры.

3.6 Оптимизация взвешивания усреднения между моделями градиентного бустинга и экстремальных деревьев

Последним и наиболее критичным параметром являлось взвешивание усреднения между двумя промежуточными моделями. Я решил, что лучший способ - попробовать ряд коэффициентов

Взвешивание усреднения производиться по формуле 2

$$y = \propto * y1 + (1 - \propto) * y2 \quad [2]$$

Где у – усредненный результат двух моделей;

$\propto$ – Сглаживающий фактор;

у1– результат модели градиентного бустинга;

у2 – результат модели экстра деревьев.

На рисунке 12 показано взвешивание усреднения результатов моделей градиентного бустинга и экстра деревьев. При значении сглаживающего фактора 0 показана точность модели градиентного бустинга а при значении 1 точность модели экстра деревьев. Точность усредненной модели достигла при значении сглаживающего фактора 0.7.

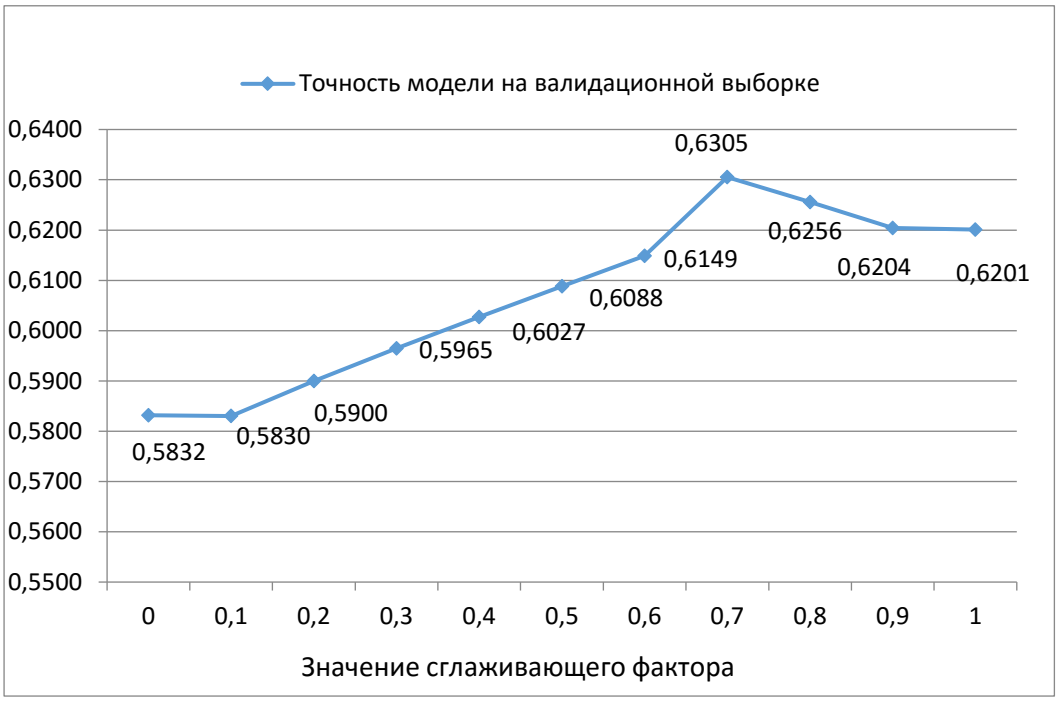

Рисунок 12 – Влияние изменения значения сглаживающего фактора усредненную модель.

Таким образом взвешивание усреднения моделей полностью обоснованно, потому что точность усредненной модели превышает каждой модели по отдельности.

3.7 Выводы по главе 3

Нами была поставлена цель – спрогнозировать время испытаний автомобильных сборок на примере компании Мерседес.

Была создана модель машинного обучения с использованием методов

– Экстремальных деревьев(Extra Tree)
– Градиентного бустинга (XGB)

Модель подошла к порогу конкурентной точности на тестовом наборе со значением коэфицента детерминации 0.55364.

4 КОММЕРЦИАЛИЗАЦИЯ ПРОЕКТА

4.1 Дорожная карта коммерциализации проекта

Дорожная карта – это детальный пошаговый план развития проекта, выработанный с учетом специфик рынка и имеющихся технологий. Хорошо составленная «дорожная карта» может помочь предсказать линии развития проекта и подобрать более успешную стратегию. Подобным образом, «дорожная карта» считается мощнейшим инструментом стратегического развития, планирования и принятия управленческих решений.

Составление «дорожной карты» — значимый период в разработке инновационного продукта. Для формирования «дорожной карты» необходимо осуществить детальный анализ рынка, исследовать технологии, дать оценку продукция, учитывать характерные черты отрасли.

Дорожная карта – это наглядное представление пошагового сценария развития определенного объекта. Процесс формирования дорожной карты:

коллективная ревизия имеющегося потенциала развития;

обнаружение возможностей роста;

обнаружение рисков;

выявление потребности в ресурсном обеспечении.

Дорожное картирование объединяет между собой виденье, стратегию и план развития объекта и создаст ключевые шаги данного процесса во времени согласно принципу «прошлое – настоящее – будущее». Оно основывается на получение экспертной информации о продукте, технологии, отрасли и т. д., позволяющей предсказывать варианты их предстоящего состояния.

4.1.1 Планирование стратегии

4.1.2 Основные цели проекта

Каждое производство, начиная с выпуска туалетной бумаги и заканчивая постройкой кораблей, имеет необходимость в некотором комплекте услуг. Подобным образом наша организация осуществляет свою работу в наиболее активной сфере – сфере услуг.

Ключевая задача проекта состоит в разработке веб-представительства компании, направленного на предоставление услуг по оптимизации процесса тестирования автомобилей.

Наличие веб-представительства даст компании следующие преимущества и решение таких задач, как:

организация получения стабильной прибыли;

формирование и продвижение имиджа компании;

увеличение спроса на предоставляемую услугу;

улучшение системы связей с общественностью;

обеспечение потребителей, партнеров, рекламных агентов полной и актуальной информацией о товаре и фирме;

обеспечение информационной поддержки потребителей посредством обратной связи;

расширение каналов сбыта предприятия.

Тем не менее основной областью деятельности планируется разработка, сбыт и обслуживание согласно предоставлению услуги, кроме того в последующем создании и доработке имеющегося программного решения на основе новейших информационных технологий. Данная сфера считается довольно молодой для отечественного рынка и по этой причине большая часть компаний ощущают недостаток в профессиональных, качественных услугах. Необходимость в данной услуге весьма велика. Все это дает возможность обеспечить требуемые решения для продуктивного функционирования различного рода фирм.

4.1.3 Источники доходов

Доходность предприятия подразумевает распространение (продажу) предоставляемых услуг, а также размещение интернет рекламы, схожей тематики (услуги по оптимизации производственной цепи, дополнительное ПО и тд.), на разработанном интернет ресурсе. Планируется что реклама будет осуществляться оп модели CPM – цена, устанавливаемая за тысячу показов. Доходность предприятия отображена в таблице 4.

Таблица 4 – Доходность веб-представительства компании

Период	Вид услуги	Объем реализации в месяц, шт.	Стоимость, руб.	Прибыль от реализации, руб.
1-3 месяц	Предоставление услуги	5-15 шт.	15 000	75 000 – 225 000
	Реклама	1 шт.	350 CPM (показов – 2 000)	700
4-6 месяц	Предоставление услуги	20-35 шт.	25 000	500 000 – 875 000
	Реклама	1-3 шт.	700 CPM (показов – 10 000)	7 000

4.1.4 Оценка использования Web-технологий в бизнесе

4.1.4.1 Целевая аудитория, конкурентная среда и потенциальные партнеры

В целевой аудитории веб-сайта можно выделить следующие группы:

Предоставляемая услуга рассчитана на уменьшение расходов на тестирование автомобилей предприятием Mersedes

Целевая аудитория: являются предприятия ведущие бизнес в аналогичной отрасли.

География: направлено на зарубежную компанию Mersedes, в дальнейшем привлечение потребителей в Российской Федерации и иных странах, и государствах;

Средний доход ЦА: любое крупное предприятие, которое в полной мере может позволить себе приобретение данной услуги;

Сфера деятельности: предоставление крупным корпорация (фирмам) услуги по оценке времени испытаний автомобильных сборок.

Интересы: Желание оптимизировать процесс тестирования автомобилей, тем самым снизив издержки в данном бизнес-процессе.

В настоящее время количество Интернет-ресурсов, реализующих те или иные услуги или программные средства достаточно велико. Однако Web-сайтов, предлагающих целенаправленную услугу по решению данной проблемы не так много, особенно те, кто предлагает эксклюзивные решения, чем и будет являться данная услуга. В связи со всем вышесказанным следует сделать вывод, что конкуренцию могут составлять только Программисты (Freelance) или же участники конкурсных разработок.

Потенциальными партнерами следует выделить организации, которые могут предоставить рекламу или размещение данной услуги на своем Web-ресурсе.

4.1.4.2 SWOT-анализ

Для того чтобы определить внешние и внутренние факторы, влияющие на возможности работы в Интернете и формирование интернет-стратегии, следует провести SWOT-анализ, который продемонстрирован в таблице 5.

Таблица 5 – SWOT-анализ

Сильные стороны	Слабые стороны
Низкая себестоимость проекта, так	Слабый маркетинг;

как он уже узкоспециализирован; Использование инновационных технологий; Сопровождение проекта, ведение и поддержка на всех этапах реализации.	Узкая направленность продукта; Отдельная оплата доп. услуг (личный менеджер, обучение персонала, тех. поддержка и др.)
Возможности	Угрозы
Новые технологии; Дополнительные услуги; Сотрудничество с другими (схожими) организациями; Увеличение рекламы;	Конкуренция в отдельно взятых конкурсах, программисты freelance; Зависимость предприятия; Лицензионный барьер, новые законодательные акты;

4.2 Создание сайта - эффективного инструмента маркетинга

4.2.1 Варианты доменного имени для сайта

Как придумать доменное имя для своего Интернет-ресурса?

Выбор домена для своего интернет-магазина аналогичен со схемой придумывания названия обычного магазина. Основными методами для этого являются:

«мозговой» штурм

оформление заказа на неймингна одной или нескольких бирж фриланса

Для успешной работы интернет-представительства компании, доменное имя должно соответствовать некоторым критериям:

быть созвучным тематике бизнеса, либо продаваемых товаров;

легкость написания, произнесения, и запоминания;

быть свободным, т.е. незарегистрированным кем-нибудь другим.

В настоящее время регистраторы доменных имен предлагают на выбор более 740 различных доменных зон, из которых фактически для бизнеса подойдет не более десятка. Топ-4 самых популярных зон – это .ru, .com, .net, org. В таблице 6 представлены варианты доменных имен, выделены их достоинства и недостатки.

Таблица 6 – достоинства и недостатки доменных имен

Домен	Достоинства	Недостатки
CarPredict.com/.ru	Незарегистрированный домен; Созвучно с тематикой; Частично отражает суть услуги и сайта; Просто запомнить; Популярная зона.	Частично понятна суть предлагаемой услуги и наполнения сайта;
time.com/.ru	Созвучно с тематикой; Частично отражает суть услуги и сайта; Просто запомнить; Популярная зона.	Частично понятна суть предлагаемой услуги и наполнения сайта; Домен уже зарегистрирован
Time-CarPredict.com/.ru	Незарегистрированный домен; Созвучно с тематикой; Понятна суть предлагаемой услуги и сайта в целом; Популярная зона.	Сложность в запоминании и написания ссылки сайта.

Окончание таблицы 6

Домен	Достоинства	Недостатки
CRPD.com/.ru	Незарегистрированный домен; Краткое запоминающееся название; Популярная зона.	Непонятен смысл сайта и предлагаемой услуги; В полной мере не отражает сущность наполнения сайта, предлагаемой услуги.

4.2.2 Тип сайта для веб-представительства компании

Landing page – это «легкий» сайт, созданный для привлечения целевой аудитории к товарам, услугам или акциям. Обычно на целевую страницу попадают благодаря переходу с контекстной рекламы или информации поисковиков. На подобных одностраничных сайтах расположена необходимая для посетителя информация в такой форме, чтобы он максимально сфокусировался на ней. Более того, правильный лендинг направлен на стимулирование желания совершить полезное действие: регистрация на сайте, оформление заказа, звонок в офис компании, подписка на рассылку. Благодаря такой направленности landing page обеспечивает повышение конверсии до 30% и более. Как правило, landing page имеют привлекательный и в меру лаконичный дизайн. Все делается для того, чтобы на странице отсутствовали факторы, отвлекающие от ее содержания.

Преимущества успешного лендинга:

ориентируясь на конкретную целевую аудиторию при правильной раскрутке и рекламе, конверсия landing page будет намного больше, чем у обычных сайтов;

благодаря простоте создания страницы она может быть готова к работе и запущена за несколько часов, а изменение информации на ней происходит в считанные минуты;

посадочные страницы обычно быстро загружаются, даже на устройствах со слабым интернетом, посетителю не надо долго ждать;

landing page – это весьма действенный и результативный инструмент, ведь если даже посетитель сайта ничего не приобретет или не закажет, велика вероятность, что он оставит свои данные. Таким образом, сформируется база потенциальных клиентов, которым в дальнейшем можно напоминать о себе по средствам e-mail рассылки;

при помощи landing page можно успешно повысить эффект от контекстной рекламы

лендинг пейдж позволяет оценить и проанализировать объемы и целесообразность интернет-продаж

помогает увеличить продажи при некачественном основном сайте

низкая стоимость разработки.

4.2.3 Информационное наполнение сайта

Тип и формат представления информации: текст и картинки обозначающие программное обеспечение;

Структурирование информации:

логотип и заголовок;

демонстрация услуги;

преимущества данного решения, в дальнейшем возможные акции;

описание оффера;

коммуникация.

Форма подачи информации:

Призыв используется в описание заголовка услуги. Призывает пользователя к действию нажать на кнопку заказа данной услуги или получения консультации по данному решению.

Аргументация используется в описание преимуществ программного решения, а также в описание оффера. Из названия следует, что пост-аргументация приводит доказательства определенной точки зрения. Его отличительная черта – вопрос «почему?», с которого начинается заголовок (например, почему предприятия должны радоваться появлению данного программного продукта: четыре причины).

Наполнение, расширение и актуализация информации. Landing page будет разбит на блоки:

Главный экран – его функция – произвести нужное впечатление на человека, информировать о том, куда он попал, мотивировать остаться и проскроллить страницу вниз.

Рассказ о проекте – подробное описание продукта или услуги:

как устроен,

как работает,

на кого ориентирован,

сколько стоит.

Рассказ о проекте – невозможно проигнорировать. Прежде чем объяснять выгоды или призывать совершить действие, необходимо убедиться, что человек понял, что именно вы предлагаете.

Понятные выгоды – этот раздел нужен, чтобы объяснить, чем вы отличаетесь от конкурентов. На большинстве рынков конкуренция высокая, поэтому необходимы доводы, почему человек должен выбрать вас.

Блоки доверия – эта группа блоков помогает сформировать кредит доверия. Отзывы, истории успеха, гарантии и сертификаты, партнеры и даже телефон и

адрес офиса помогут развиртуализировать проект, показать, что он реальный и ему можно доверять.

Целевое действие – Бизнесу нужны клиенты, поэтому на лендинге должны быть блоки, которые будут генерировать лиды: формы заказа, подписки, обратной связи или телефон.

Дизайн сайта (обложка), наполнение:

Главный экран сайта – первое впечатление от компании. Есть всего несколько секунд, чтобы убедить пользователя остаться на странице.

Набор инструментов для этого небольшой: заголовок, подзаголовок, кнопка или форма, логотип, фон или изображение на фоне, меню, стрелка вниз.

Заголовок и подзаголовок – сделайте оффер – вдохновляющую фразу, которая передает суть проекта. Как правило, заголовок более эмоциональный, подзаголовок раскрывает смысл.

Форма или кнопка – для тех, кто сразу заинтересовался или зашел повторно, можно сразу на обложке добавить целевое действие.

Фон обложки – хорошая фотография, атмосферное видео, просто цвет, градиент или иллюстрация. Стоит обратить внимание на сочетание фона с текстом: фотография может быть удачной сама по себе, но если она неоднородная, пестрая, то она будет плохо работать с текстом. Видео нужно снимать, во-первых, плавно, во-вторых, лучше брать увеличенный фокус, чтобы все объекты были крупноваты.

Логотип – компании или продукта можно расположить как на самой обложке, так и в меню.

Стрелка вниз – Не обязательный элемент обложки. Стоит ориентироваться на аудиторию – если она консервативная, то стоит добавить. Новое поколение привыкло к скроллу, но кто-то может застопориться.

Меню – Также не обязательный элемент, но если он нужен для навигации, то выделить основные смысловые секции на странице, к которым нужен быстрый доступ.

На рисунке 1 продемонстрирована главная страница Web-ресурса. Предлагаемая услуга отображена на рисунке 2.

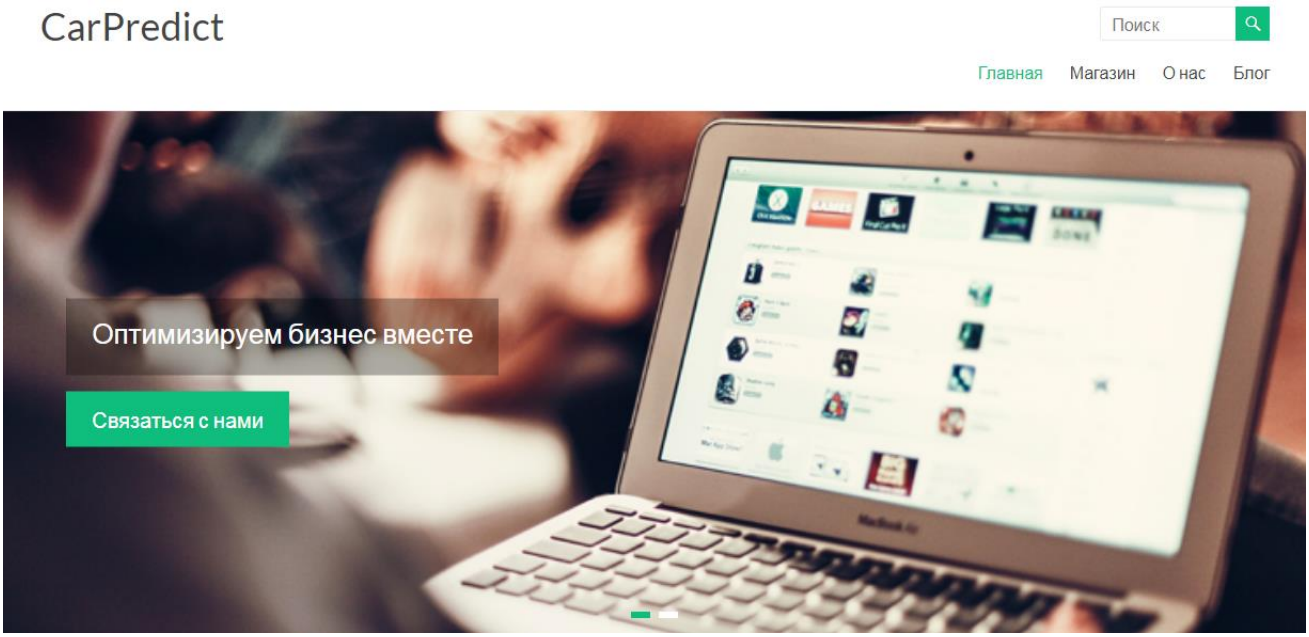

Рисунок 1 – Главная страница Landing page

Прогнозирование времени тестирования автомобилей

15000.00$

Благодаря данной программе вы можете оптимизировать процесс тестирования автотранспорта что позволит снизить издержки в данной области.

В корзину

Рисунок 2 – Предлагаемая услуга

4.3 Инструменты работы с аудиторией сайта

Анализ поведения пользователей на сайте – владельцы ресурса могут следить за посещаемостью сервера, за наиболее популярными маршрутами по сайту, точками входа и выхода посетителей, временем, проведенным на каждой из страниц и т.д. Данная информация используется и для определения эффективности рекламных направлений, и для оптимизации структуры и навигации сайта. Получать подобные данные можно с помощью анализатора логов сайта или продвинутых счетчиков.

Консультации – с помощью интернет-технологий можно эффективно осуществлять информационную поддержку своих клиентов. Специалисты компании

с помощью on-line конференций, чата или по e-mail могут отвечать на вопросы, давать консультации. В случае с конференцией это будет не столь оперативно (хотя и конференции могут проводиться в реальном режиме времени), но наглядно и информативно. Конференции имеют удобную древовидную структуру, а отсутствие необходимости отвечать сразу позволяет более тщательно подготовить ответ.

Чат – в дальнейшем планируется разработка чата. Он дает максимальную оперативность, ту же, что и телефонная линия, но при этом не надо платить за международные переговоры, а специалист службы поддержки может одновременно отвечать сразу на несколько вопросов. Самым же распространенным способом поддержки пользователей остаются консультации посредством электронной почты.

Патчи, драйвера и обновления программ – продавцы программного обеспечения, помимо консультаций и инструкций, посредством Интернета могут распространять как непосредственно свою продукцию, так и патчи и обновления к ней. А производители высокотехнологического оборудования могут выкладывать на сайте для скачивания последние версии драйверов устройств.

4.4 Мониторинг сайта

В таблице 7 отображена полная информация о возможностях сайта.

Таблица 7 – Возможности сайта

Наименование	CarPredict
Информационное наполнение	Информации четко структурирована (в дальнейшем планируется добавлять новости и обновления в блог), используются различные форматы представления информации,

	тех. поддержка, адекватная и структурированная информация сайта, имеется расстановка информационных акцентов.
Функциональность	Представление товара и формирование заказа осуществляется по нажатию кнопки и переходу на вспомогательный экраны, а также окно формирования заказа, присутствует связь при помощи e-mail или телефона (в дальнейшем предусмотрена разработка чат-бота).
Usability	Сайт эргономичен и удобен в использовании (присутствует простая и эффективная навигация, имеется карта местоположения, привычный вид полей и кнопок.
Дизайн	Дизайн дополняет и усиливает заложенную в сайт информацию и функционал, простота использования сайта благодаря легкому дизайну, возможность изменения дизайн-решений , уникальность и запоминаемость.

Окончание таблицы 7

Наименование	CarPredict
Техническая реализация	Сайт написан при помощи движка WordPress.
Маркетинг	На сайте присутствуют адреса, ссылки на сайт, средства сбора информации о посетителях сайта, посещаемость и поведенческая линия на сайте, работа с аудиторией сайта.

4.5 Продвижение и ценовая политика сайта

4.5.1 Медиаплан, ценовая политика

Медиапланирование – это планирование каналов и способов рекламы для составления медиаплана на основе прогнозов и полученных результатов.

Медиаплан для первого рекламного мероприятия по продвижению, созданного Web-ресурса, должен выполнять следующие условия:

Бюджет – 10 000-30 000 $ (по нынешнему курсу рубль-$ = 64,60);

Время рекламной компании – 4 недели;

Задача рекламной компании – привлечение посетителей (раскрутка нового ресурса).

Для рекламы конечно же будут использоваться самые распространенные и популярные рекламные площадки такие как Вконтакте и Яндекс. В таблице 8 продемонстрирована полная информация по выбору той или иной площадки.

Для начала, чтобы раскрутить бренд следует максимизировать сиюминутную прибыль. Иначе говоря - извлечь как можно больше денег из каждой продажи предоставляемой услуги, даже если это сокращает количество потенциальных покупателей. В итоге, будет меньше клиентов, но и количество проблем по их

обслуживанию также сократится. И, кроме того, каждый из клиентов принесет больший доход.

Расчет будет производиться по общим издержкам – сумма постоянных и переменных издержек.

"Снятие сливок". Предлагая новую революционную услугу, стоит изначально установить на нее высокую цену, так как тот, кто чувствителен к нововведениям – нечувствителен к цене. Затем снижаем цену и "снимаем сливки" со следующего слоя покупателей и так далее. В конечном счете, цена падает под воздействием того, что товар укрепляет свои позиции на рынке и конкуренты снижают цены. Так же стоит задуматься о скидках на повторное приобретение услуги если в таковой нуждаются.

Таблица 8 – Медиаплан

Рекламные каналы	Дополнительная информация	Посыл	Формат	Общая стоимость рекламы, руб.	Число публикаций	CPT, руб.	Частота	Бюджет
Контекстная реклама в Яндексе	Только горячие запросы	Инновационная разработка	Спец. размещение	120 000	1	1 500	1	120 000
РСЯ	Публикация рекламных постов	Инновационная разработка	Графические баннеры	450 000	3	40	4	1 350 000
Группа Вконтакте	Публикация рекламных постов	Инновационная разработка	Промо-пост	10 000	2	833	1,2	20 000
Таргетированная реклама Вконтакте	Публикация рекламных постов	Инновационная разработка	Лид-форма	100 000	3	1 500	1,2	300 000
Mail.Ru Group	Рекламный баннер	Инновационная разработка	Графические баннеры	55 000	2	900	1	110 000
Итого:				735 000	11			1 900 000

7

4.6 Выводы по главе 4

Составление «дорожной карты» — важный этап в создании инновационного продукта. Благодаря составлению которой, было спланировано веб-представительство будущей компании, деятельность которого направлена на предоставление услуг по предотвращению и сокращению технологических сбоев производственной линии на предприятии.

Нами была составлена таблица примерной доходности веб-ресурса. Рассмотрена целевая аудитория и проведен SWOT-анализ для определения внешних и внутренних факторов, влияющих на возможности работы в Интернете и формировании интернет стратегии.

Подходящим вариантом типизации сайта был выбран Landing page с доменным именем Prediction.com/.ru. После чего, продумано информационное наполнение, продемонстрирован первоначальный предполагаемый вид и раскрыты возможности сайта.

В дальнейшем планируется разработать инструменты работы с аудиторией, такие как:

- анализ поведения пользователей на сайте;
- консультации;
- чат;
- патчи, драйвера и обновления программ.

Разработан медиаплан и ценовая политик

5 ЗАКЛЮЧЕНИЕ

1. Мы рассмотрели процесс испытания автомобильных сборок. Это производственный процесс, в котором производится контроль качества выпускаемой продукции. Были рассмотрены факторы влияющие на время испытаний автомобильных сборок и методы снижения данного времени. Рассмотрены наиболее важные правила, которые необходимо учитывать при создании системы контроля качества. Исходя из исследования предметной области, мы пришли к выводу о том, что для решения задачи прогнозирования времени испытания автомобильных сборок эффективно использовать методы машинного обучения.

2. В выпускной квалификационной работе были изучены регрессионные методы прогнозирования времени испытания автомобильных сборок. Рассмотрены популярные алгоритмы, которые используются в машинном обучении для решения данных проблем, такие как линейная регрессия, случайный лес (Random forest), экстра-деревья (Extra-trees), градиентный бустинг (XGBoost). Проанализирована используемая метрика качества.

3. Была создана модель машинного обучения с использованием методов

- Экстремальных деревьев(Extra Tree)
- Градиентного бустинга (XGB)

Модель подошла к порогу конкурентной точности со значением коэфицента определения 0.55364.

4. Разработана коммерциализация проекта по этапам. Изначально была разработана дорожная карта коммерциализации данного проекта, которая подразумевает наглядное представление пошагового сценария развития, в которую входит – планирование стратегии, исходя из задач, для решения поставленной цели, описаны источники доходов по видам предоставляемых услуг и их стоимость в рублях. Проведена оценка потенциальных возможностей Интернета для бизнеса, в которой были рассмотрены: целевая аудитория, конкурентная среда и потенциальные партнеры. Также продемонстрирована таблица SWOT-анализа. По потенциальным возможностям Интернета было выбрано создать сайт по предоставлению

услуги прогнозирования сбоев технологических линий другим компаниям. Для решения данной задачи первостепенным было принято решение выбора доменного имени для сайта, а также был выбран тип и информационное наполнение сайта. Следующим шагом был выбор инструментов для работы с аудиторией сайта. В табличном виде представлен мониторинг сайта. Описано продвижение и ценовая политика сайта. Разработан медиаплан, также продемонстрированный в табличном виде.

6 БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ankita Mangal, Elizabeth A. Holm, "Applied Machine Learning to Predict Stress Hotspots I: Face Centered Cubic Materials", International Journal of Plasticity, 2018.
2. Ankita Mangal, Elizabeth A. Holm, "Applied Machine Learning to Predict Stress Hotspots II: Hexagonal close packed materials", International Journal of Plasticity, 2018.
3. Felix Reinhart, Sebastian von Enzberg, Arno Kühn, Roman Dumitrescu, Machine Learning for Cyber Physical Systems, vol. 11, pp. 25, 2017.
4. H. B. McMahan, G. Holt, D. Sculley, M. Young,
5. D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov,
6. D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica, “Ad Click Prediction: a View from the Trenches,” 2013.
7. B. Zenko, “Is Combining Classifiers Better than Selecting the Best One?” Machine Learning, vol. 54, no. 3, pp. 255–273, 2004.
8. O´ . Fontenla-Romero, B. Guijarro-Berdin˜as,
9. D. Martinez-Rego, B. Pe´rez-Sa´nchez, and D. Peteiro- Barral, “Online machine learning,” Efficiency and Scalability Methods for Computational Intellect, pp. 27–54, 2013.

    A. Y. Ng, “Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance,” in Proceedings of the Twenty-first International Conference on Machine Learning, ser. ICML ’04. New York, NY, USA: ACM, 2004, p. 78. [Online]. Available: http://doi.acm. org/10.1145/1015330.1015435

10. J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich, “Collaborative Email-Spam Filtering with the Hashing-Trick,” Conference on Email and Anti-Spam, pp. 1–4, 2009. [Online]. Available: http://cran.fhcrc.org/web/packages/xgboost/ vignettes/xgboost.pdf
11. J. Friedman, T. Hastie, R. Tibshirani, and Others, “Ad- ditive logistic regres-sion: a statistical view of boosting (with discussion and a rejoinder by the authors),” The annals of statistics, vol. 28, no. 2, pp. 337–407, 2000.
12. J. H. Friedman, “Greedy function approximation: a gradient boosting ma-chine,” Annals of statistics, pp. 1189–1232, 2001.
13. B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochimica et Biophysica Acta (BBA)-Protein Struc- ture, vol. 405, no. 2, pp. 442–451, 1975.