ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Data Science»

Слушатель Степанова Василиса Валерьевна

Москва, 2023

Содержание

Введение

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента). На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ

«Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана). Актуальность: Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита. Для достижения определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов.

1. Аналитическая часть

1.1 Постановка задачи

Для исследовательской работы были даны 2 файла: X_bp.xlsx (с данными о параметрах, состоящий из 1023 строк и 10 столбцов данных) и X_nup.xlsx (данными нашивок, состоящий из 1040 строк и 3 столбцов данных).

Для разработки моделей по прогнозу модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель нужно объединить 2 файла. Объединение по типу INNER, поэтому часть информации (17 строк таблицы X_nup.xlsx) не имеет соответствующих строк в таблице X_bp.xlsx и будет удалена.

Также необходимо провести разведочный анализ данных, нарисовать гистограммы распределения каждой из переменной, диаграммы boxplot (ящик с усами), попарные графики рассеяния точек.

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; сделать предобработку: удалить шумы и выбросы, сделать нормализацию и стандартизацию.

Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз соотношения «матрица-наполнитель». Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

1.2 Описание используемых методов

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и традиционно это задача

регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её, поэтому для наилучшего решения были исследованы (и некоторые из них применены) следующие методы:

− линейная регрессия (Linear regression);

− лассо регрессия (Lasso);

− гребневая регрессия (Ridge);

− эластичная регрессия (ElasticNet) ;

− градиентный бустинг(GradientBoostingRegressor);

− К-ближайших соседей (KNeighborsRegressor);

− дерево решений (DecisionTreeRegressor);

− случайный лес (RandomForest);

− градиентный бустинг (AdaBoostRegressor);

− стохастический градиентный спуск (SGDRegressor);

− метод опорных векторов (Support Vector Regression);

− многослойный перцептрон.

Линейная регрессия (Linear regression) — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R2, или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R2 к единице, тем лучше.

Достоинства метода: быстр и прост в реализации; легко интерпретируем, имеет меньшую сложность по сравнению с другими алгоритмами.

Недостатки метода: моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

Чтобы улучшить Линейную модель путем обмена некоторой этой дисперсии с предвзятостью, чтобы уменьшить нашу общую ошибку. Это происходит при помощи регуляризации, в которой модифицируется функция стоимости, чтобы ограничить значения коэффициентов. Это позволяет изменить чрезмерную дисперсию на некоторое смещение, потенциально уменьшая общую ошибку.

Лассо регрессия (Lasso) – это линейная модель, которая оценивает разреженные коэффициенты. Это простой метод, позволяющий уменьшить сложность модели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Это даёт более устойчивое решение. В регрессии лассо добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квадратичного смещения, используется смещение абсолютного значения. Лассо регрессия хорошо прогнозирует модели временных рядов на основе регрессии, таким как авторегрессии.

Достоинства метода: легко полностью избавляется от шумов в данных; быстро работает; не очень энергоёмко; способно полностью убрать признак из датасета; доступно обнуляет значения коэффициентов.

Недостатки метода: часто страдает качество прогнозирования; выдаёт ложное срабатывание результата; случайным образом выбирает одну из коллинеарных переменных; не оценивает правильность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Лассо-регрессию следует использовать, когда есть несколько характеристик с высокой предсказательной способностью, а остальные бесполезны. Она обнуляет бесполезные характеристики и оставляет только подмножество переменных.

Гребневая регрессия (Ridge) – это регрессия, которая добавляет дополнительный штраф к функции стоимости, но вместо этого суммирует квадраты значений коэффициентов (норма L-2) и умножает их на некоторую постоянную лямбду. По сравнению с Лассо этот штраф регуляризации уменьшит значения коэффициентов, но не сможет принудительно установить коэффициент равным 0. Это ограничивает использование регрессии гребня в отношении выбора признаков. Однако, когда p> n, он способен выбрать более n релевантных предикторов, если необходимо, в отличие от Лассо. Он также выберет группы коллинеарных элементов, которые его изобретатели назвали «эффектом группировки».

Как и в случае с Лассо, мы можем варьировать лямбду, чтобы получить модели с различными уровнями регуляризации, где лямбда = 0 соответствует OLS, а лямбда приближается к бесконечности, что соответствует постоянной функции.

Анализ регрессии Лассо, так и Риджа показазывает, что ни один метод не всегда лучше, чем другой; нужно попробовать оба метода, чтобы определить, какой использовать.

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

Ридж-регрессию лучше применять, когда предсказательная способность набора данных распределена между различными характеристиками. Ридж-регрессия не обнуляет характеристики, которые могут быть полезны при составлении прогнозов, а просто уменьшает вес большинства переменных в модели.

Эластичная сеть (ElasticNet) – это регрессия, которая включает в себя термины регуляризации как L-1, так и L-2. Это дает преимущества регрессии Лассо и Риджа. Было установлено, что он обладает предсказательной способностью лучше, чем у Лассо, хотя все еще выполняет выбор функций. Поэтому получается лучшее из обоих методов, выполняя выбор функции Лассо с выбором группы объектов Ridge.

Elastic Net поставляется с дополнительными издержками на определение двух лямбда-значений для оптимальных решений.
Компромисс смещения дисперсии — это компромисс между сложной и простой моделью, в которой промежуточная сложность, вероятно, является наилучшей.

Лассо, Ридж-регрессия и Эластичная сеть — это модификации обычной линейной регрессии наименьших квадратов, которые используют дополнительные штрафные члены в функции стоимости, чтобы сохранить значения коэффициента небольшими и упростить модель.

Лассо полезно для выбора функций, когда наш набор данных имеет функции с плохой предсказательной силой.

Регрессия гребня полезна для группового эффекта, при котором коллинеарные элементы могут быть выбраны вместе.

Elastic Net сочетает в себе регрессию Лассо и Риджа, что потенциально приводит к модели, которая является простой и прогнозирующей.

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

Градиентный бустинг (Gradient Boosting) — это ансамбль деревьев решений, обученный с использованием градиентного бустинга. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Основная идея градиентного бустинга: строятся последовательно несколько базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов. Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, или это может привести к переобучению, наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.
Метод ближайших соседей - К-ближайших соседей (kNN - k Nearest Neighbours) ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. Алгоритм находит расстояния между запросом и всеми

примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Достоинства метода: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоемкость.

Дерево решений (DecisionTreeRegressor) – метод автоматического анализа больших массивов данных. Это инструмент принятия решений, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево принятия решений - эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область. Регрессия дерева решений отслеживает особенности объекта и обучает модель в структуре дерева прогнозированию данных в будущем для получения значимого непрерывного вывода. Дерево решений один из вариантов решения

регрессионной задачи, в случае если зависимость в данных не имеет очевидной корреляции.

Достоинства метода: помогают визуализировать процесс принятия решения и сделать правильный выбор в ситуациях, когда результаты одного решения влияют на результаты следующих решений, создаются по понятным правилам; просты в применении и интерпретации; заполняют пропуски в данных наиболее вероятным решением; работают с разными переменными; выделяют наиболее важные поля для прогнозирования.

Недостатки метода: ошибаются при классификации с большим количеством классов и небольшой обучающей выборкой; имеют нестабильный процесс (изменение в одном узле может привести к построению совсем другого дерева); имеет затратные вычисления; необходимо обращать внимание на размер; ограниченное число вариантов решения проблемы.

Случайный лес (RandomForest) — это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов. Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать вместе.

Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую параллелизуемость и масштабируемость.

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недообучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

Градиентный бустинг (AdaBoost) – это алгоритм, который работает по принципу перевзвешивания результатов. Есть деревья решений, а ансамбль из них это градиентный бустинг, задача решается с помощью градиентсного

спуска. Алгоритм AdaBoost учится на ошибках, больше концентрируясь на сложных участках, с которыми от столкнулся в процессе предыдущей итерации обучения. На каждой итерации дается вес алгоритмам. Каждый новый алгоритм корректирует ошибки предыдущих до получения хорошего результата. Все прогнозы объединяются с помощью голосования для получения окончательного прогноза.

Достоинства метода:

AdaBoost легко реализовать, достаточно класса моделей и их количества.

Он итеративно исправляет ошибки слабого классификатора и повышает точность путем объединения слабых учащихся.

Можно использовать многие базовые классификаторы с AdaBoost.

AdaBoost не склонен к переоснащению.

Недостатки метода:

AdaBoost чувствителен к шумным данным.

AdaBoost обучается дольше линейной регрессии, классификация дольше чем при использовании логистической регрессии.

На AdaBoost сильно влияют отклонения, так как он пытается идеально подогнать каждую точку.

AdaBoost работает медленнее и чуть хуже, чем XGBoost. Но легче в понимании.

Стохастический градиентный спуск (SGDRegressor) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь. Этот подход подразумевает корректировку весов нейронной сети, используя аппроксимацию градиента функционала, вычисленную только на одном случайном обучающем примере из выборки.

Достоинства метода: эффективен; прост в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках.

Недостатки метода: требует ряд гиперпараметров; чувствителен к масштабированию функций; может не сходиться или сходиться слишком

медленно; функционал многоэкстремален; процесс может "застрять" в одном из локальных минимумов; возможно переобучение.

Метод опорных векторов (Support Vector Regression) – этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших датасетах. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использованием схожих алгоритмов для анализа данных и распознавания шаблонов. Учитывая обучающую выборку, где алгоритм помечает каждый объект, как принадлежащий к одной из двух категорий, строит модель, которая определяет новые наблюдения в одну из Категорий. Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв.

Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию. Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

Недостатки метода: неустойчивость к шуму, поэтому в работе была проведена тщательнейшая работа с выбросами, иначе в обучающих данных шумы становятся опорными объектами-нарушителями и напрямую влияют на построение разделяющей гиперплоскости; для больших наборов данных

требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать, поэтому были рассмотрены и другие методы.

Многослойный персептрон (MLPRegressor) — это алгоритм обучения с учителем, который изучает функцию $f(\cdot):Rm \rightarrow Ro$ обучением на наборе данных, где m — количество измерений для ввода и o- количество размеров для вывода. Это искусственная нейронная сеть, имеющая 3 или более слоёв персептронов. Эти слои - один входной слой, 1 или более скрытых слоёв и один выходной слой персептронов.

Достоинства метода: построение сложных разделяющих поверхностей; возможность осуществления любого отображения входных векторов в выходные; легко обобщает входные данные; не требует распределения входных векторов; изучает нелинейные модели.

Недостатки метода: имеет невыпуклую функцию потерь; разные инициализации случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

Используемые метрики качества моделей:

R2 (коэффициент детерминации) измеряет долю дисперсии, объяснённую моделью, в общей дисперсии целевой переменной.

Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

MSE (Mean Squared Error) (средняя квадратичная ошибка) принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

1.3. Разведочный анализ данных

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Необработанные данные могут содержать искажения и пропущенные значения и способны привести к неверным результатам по итогам моделирования. Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

Рисунок 1 - Описательная статистика датасета

Цель разведочного анализа - получение первоначальных представлений о характерах распределений переменных исходного набора данных, формирование оценки качества исходных данных (наличие пропусков, выбросов), выявление характера взаимосвязи между переменными с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

Рисунок 2 - Проверка датасета на наличие дубликатов

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной; диаграммы boxplot (ящика с усами); попарные графики рассеяния точек; тепловая карта; описательная статистика для каждой переменной; анализ и полное исключение выбросов; проверка наличия пропусков и дубликатов; корреляция Кендалла и Пирсона.

Рисунок 3

Рисунок 4

Для удаления выбросов используются методы трех сигм и межквартильного расстояния. В данном случае удалим способом межквартильного расстояния для максимальной чистоты, так как используем методы чувствительные к выбросам.

Рисунок 6 - Тепловая карта с корреляцией данных

Тепловая карта показывает практически отсутствие корреляции между признаками и целевыми переменными.

Рисунок 7 - Гистограммы распределения

Гистограммы показывают нормальное распределение, за исключением признака Угол нашивки, который имеет всего два значения 0 и 90 градусов.

Рисунок 8 - Попарные графики рассеяния точек с выделением значений Угол нашивки

На попарных графиках распределения не видно корреляции между признаками. Единственная зависимость, которую можно отметить – это меньшая дисперсия значений Плотности нашивки при 90 градусов Угла нашивки, по сравнению со значением 0 градусов.

Максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

2. Практическая часть

2.1. Разбиение и предобработка данных

2.1.1 Для прогнозирования модуля упругости при растяжении

Признаки датасета были разделены на входные и выходные, а строки - на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 8. Описательная статистика входных признаков до и после предобработки показана на рисунке 9. Описательная статистика выходного признака показана на рисунке 10.

2.1.2 Для прогнозирования прочности при растяжении

Признаки датасета были разделены на входные и выходные, а строки - на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 11. Описательная статистика входных призна-ков до и после предобработки показана на рисунке 12. Описательная стати-стика выходного признака показана на рисунке 13.

2.1.3 Для прогнозирования соотношения матрица-наполнитель

Признаки датасета были разделены на входные и выходные, а строки - на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 14. Описательная статистика входных призна-ков до и после предобработки показана на рисунке 15. Описательная стати-стика выходного признака показана на рисунке 16.

2.2 Разработка и обучение моделей для прогнозирования модуля упругости при растяжении

Для подбора лучшей модели для этой задачи я взяла следующие моде-ли:

 LinearRegression — линейная регрессия (раздел 1.2.1);

 Ridge — гребневая регрессия (раздел 1.2.2);

 Lasso — лассо-регрессия (раздел 1.2.2);

 SVR — метод опорных векторов (раздел 1.2.3);

 KneighborsRegressor — метод ближайших соседей (раздел 1.2.4);

 DecisionTreeRegressor — деревья решений (раздел 1.2.5);

 RandomForestRegressor — случайный лес (раздел 1.2.6).

В качестве базовой модели взят DummyRegressor, возвращающий среднее значение целевого признака.

Метрики работы выбранных моделей с гиперпараметрами по умолча-нию, полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 17.

Ни одна из выбранных мной моделей не оказалась подходящей для наших данных.

Коэффициент детерминации R2 близок к 0 для линейных моделей и ме-тода опорных векторов. Значит, они не лучше базовой модели. И остальные метрики у них примерно совпадают с базовой моделью.

Гораздо хуже линейных моделей с гиперпараметрами по умолчанию отработали метод ближайших соседей и деревья решений.

Случайный лес отработал лучше, чем одно дерево решений, но хуже, чем линейные модели.

После выполнения подбора гиперпарметров по сетке с перекрестной проверкой, получили метрики, приведенные на рисунке 18.

Можно сделать вывод, что подбирая гиперпараметры, можно значи-тельно улучшить предсказание выбранной модели.

Все модели крайне плохо описывают исходные данные - не удалось до-биться положительного значения R2. Самая лучшая модель дает коэффици-ент детерминации близкий к нулю, что соответствует базовой модели.

Линейные модели совпадают с базовой моделью. Их характеристики улучшились, но не значительно.

Метод опорных векторов в процессе подбора гиперпараметры лучшим ядром выбрал линейное и отработал аналогично линейным моделям.

Метод ближайших соседей увеличением количества соседей радикально улучшил качество работы. Но его лучшие результаты все равно немного, но отстают от линейных моделей.

Деревья решений при кропотливом подборе параметров превзошли результат линейной модели. Но они не являются объясняющей зависимость моделью.

Собирая деревья в ансамбли, можно улучшать характеристики. Но подбор параметров для леса затруднен тем, что это затратный по времени процесс. По этой причине мне не удалось получить комбинацию параметров для леса, которая была был лучше дерева решений.

Поэтому в качестве лучшей модели выбираю дерево решений. На рисунке 19 приведена визуализация работы лучшей модели на тестовом множестве.

Сложно визуализировать регрессию в многомерном пространстве. Но даже на таком графике мы видим, насколько не соответствует лучшая модель исходным данным и насколько она неудачна.

Метрики работы лучшей модели на тестовом множестве и сравнение с базовой отражены на рисунке 20. Они подтверждают: полученная модель хуже базовой. Результат исследования отрицательный. Не удалось получить модели, которая могла бы оказать помощь в принятии решений специалисту предметной области.

2.4 Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель

По заданию для соотношения матрица-наполнитель необходимо по-строить нейросеть. Но для сравнения нам также понадобится базовая модель DummyRegressor, возвращающая среднее целевого признака.

2.4.1 MLPRegressor из библиотеки sklearn

Строю нейронную сеть с помощью класса MLPRegressor следующей архитектуры:

- слоев: 8;
- нейронов на каждом слое: 24;
- активационная функция: relu;
- оптимизатор: adam;
- пропорция разбиения данных на тестовые и валидационные: 30%;
- ранняя остановка, если метрики на валидационной выборке не улуч-шаются;
- количество итераций: 5000.

Нейросеть обучилась за 343мс и 33 итерации. График обучения приве-ден на рисунке 25.

2.6. Разработка приложения

Несмотря на то, что пригодных к внедрению моделей получить не уда-лось, можно разработать функционал приложения. Возможно, дальнейшие исследования позволят построить качественную модель и внедрить ее в гото-вое приложение.

В приложении необходимо реализовать следующие функции:

- выбор целевой переменной для предсказания;
- ввод входных параметров;
- проверка введенных параметров;

 загрузка сохраненной модели, получение и отображение прогноза выходных параметров.

Решено разработать веб-приложение с помощью языка Python, фрейм-ворка Flask и шаблонизатора Jinja.

Эту задачу получилось решить. Скриншоты разработанного веб-приложения приведены в приложении А.

2.7. Создание удаленного репозитория

Для данного исследования был создан удаленный репозиторий на GitHub, который находится по адресу https://github.com/e-ginger/bmstu-ds-course. На него были загружены результаты работы: исследовательский notebook, код приложения.

Заключение

В ходе выполнения данной работы мы прошли практически весь Dataflow pipeline, рассмотрели большую часть операций и задач, которые приходится выполнять специалисту по работе с данными. Этот поток операций и задач включает:

- изучение теоретических методов анализа данных и машинного обуче-ния;
- изучение основ предметной области, в которой решается задача;
- извлечение и транформацию данных. Здесь нам был предоставлен го-товый набор данных, поэтому через трудности работы с разными источни-ками и парсингом данных мы еще не соприкоснулись;
- проведение разведочного анализа данных статистическими методами;
- DataMining — извлечение признаков из датасета и их анализ;
- разделение имеющихся, в нашем случае размеченных, данных на обучающую, валидационную, тестовую выборки;
- выполнение предобработки (препроцессинга) данных для обеспече-ния корректной работы моделей;
- построение аналитического решения. Это включает выбор алгоритма решения и модели, сравнение различных моделей, подбор гиперпараметров модели;
- визуализация модели и оценка качества аналитического решения;
- сохранение моделей;

 разработка и тестирование приложения для поддержки принятия ре-шений специалистом предметной области, которое использовало бы найденную модель;

 внедрение решения и приложения в эксплуатацию. Этот блок задач мы тоже пока не затронули.

В этой работе мы имели дело не с учебными наборами данных, кото-рые дают хорошо изученные решения, а с реальной производственной зада-чей. И к сожалению, не смогли поставленную задачу решить — не получили моделей, которые бы описывали закономерности предметной области. Я проделала максимум исследований, которые в моей компетенции как начи-нающего дата-сайентиста, применила большую часть знаний, полученных в ходе прохождения курса.

Возможные причины неудачи:

 нечеткая постановка задачи, отсутствие дополнительной информации о зависимости признаков с точки зрения физики процесса. Незначимые при-знаки являются для модели шумом, и мешают найти зависимость целевых от значимых входных признаков;

 исследование предварительно обработанных данных. Возможно, на "сырых", не предобработанных данных можно было бы получить более ка-чественные модели, воспользовавшись другими методами очистки и подго-товки;

 мой недостаток знаний и опыта. Нейросети являются самым совре-менным подходам к решению такого рода задач. Они способны

находить скрытые и нелинейные зависимости в данных. Но выбор оптимальной ар-хитектуры нейросети является неочевидной задачей. Дальнейшие возможные пути решения этой задачи могли бы быть:

 углубиться в изучение нейросетей, попробовать различные архитек-туры, параметры обучения и т.д.;

 провести отбор признаков разными методами. Испробовать методы уменьшения размерности, например метод главных компонент;

 после уменьшения размерности градиентный бустинг может улуч-шить свои результаты. Так же есть большой простор для подбора гиперпа-раметров для этого метода;

 проконсультироваться у экспертов в предметной области. Возможно, они могли бы поделиться знаниями, необходимыми для решения задачи.