1. **Цель.**

Подобрать признаки и обучить модель для предугадывания свойств растворимости молекул на основе предоставленных данных.

1. **Задачи.**

Базируясь на поставленной цели, сформулированы следующие задачи:

1. Осуществить подбор потенциально полезных признаков на основе различных по своей природе представлениях молекул и таким образом сформировать исходные массивы данных для дальнейшего обучения.
2. Осуществить подбор моделей для каждого из используемых массивов данных.
3. **Использованное оборудование.**

В качестве рабочего оборудования использован персональный компьютер со следующими характеристиками:

* **Процессор:** AMD Ryzen 5 2600 Six-Core Processor, 3400 МГц, ядер: 6, логических процессоров: 12
* **Размер оперативной памяти:** 8.00 ГБ
* **Операционная система:** Майкрософт Windows 10 Домашняя
* **Язык программирования:** Python v 3.8

1. **Описание и ход решения задачи.**
   1. *Создание первичного датафрейма*

В качестве отправной точки исследования были предоставлены две таблицы: «concat.CSV» и «1000000.smi». Первая из них содержала данные по искомому признаку, на основе которого должна проверяться регрессионная модель, а также прочий список свойств. Вторая содержала пары «Кодировка» / «SMILE-строка».

В первую очередь следовало рассмотреть исходные данные и ответить на вопрос: одинаковы ли размеры исходных массивов данных? Помимо прочего, также необходимо было точно знать, однородные ли данные содержат внутри себя столбцы таблиц, нет ли среди них пустот и т. д.

Несмотря на то, что исходные данные оказались однородными, в ходе проверки с помощью средств библиотеки pandas выяснилось, что количество строк в них не соответствует друг другу. Ввиду указанных обстоятельств было принято решение создать новый массив данных, используя класс «CreateSourceDataFrame».

Указанный класс сопоставляет массиву «1000000.smi» столбец «QPlogS» из таблицы «concat.CSV», предварительно убирая дубликаты кодировок молекул и используя их в качестве индексов для слияния двух датафреймов в один. В новом датафрейме содержались колонки с кодировкой молекул, их SMILE-представлением и соответствующем ему значением десятичного логарифма растворимости вещества.

После слияния датафреймов и проверки целостности исходных данных необходимо было решить, на основе каких признаков строить модель. Для выбора исходных признаков следовало обратиться к теоретическим знаниям.

* 1. *Характеристика исходных данных*

При предоставлении исходных таблиц тестового задания не упоминалось ни то какого характера был растворитель, для которого указаны текущие величины десятичного логарифма, ни то каким образом были собраны эти данные – были ли они получены экспериментально или же с помощью вычислительных технологий? Исходя из этой некоторой недостаточности данных, для продолжения выполнения задания следовало выдвинуть следующие сильные предположения: **все данные были получены экспериментально и достоверны**; **для каждой молекулы был использован один и тот же растворитель, одинаковой чистоты и при одних и тех же внешних условиях**.

* 1. *Подбор характеристик для обучения моделей*

На растворимость вещества в том или ином растворителе влияет множество факторов, однако все их можно разделить на два «лагеря»:

1. **Термодинамичесмкие и термохимические,** основанные в основном на состоянии окружающей среды и внутренней энергии молекул (полярность, энтропия системы, внутренняя энергия и т. д.)
2. **Структурные,** основанные на природном сродстве того или иного вещества к тому или иному растворителю.

Обе совокупности факторов оказывают большое влияние на то, как хорошо будет та или иная молекула растворяться в том или ином растворителе. Рассмотрим их обе.

В первом случае мы обратимся к указанному выше предположению об идентичности условий проведения экспериментов по определению растворимости и дополним его: эксперименты проводились в **стандартных условиях**, а растворителем могла бы выступить **очищенная вода**.

Согласно второму началу термодинамики, при изобарно-изотермическом процессе **вещества могут самопроизвольно растворяться в некоем растворителе, если в результате этого процесса энергия Гиббса системы уменьшается**. В соответствии с этим законом мы можем заключить, что растворение будет тем интенсивнее, чем выше **энтропийный фактор растворения.** Иначе говоря, выдвинув предположение об изобарной и изотермической природе процесса растворения мы могли бы попытаться найти энтропию системы «вещество-вода». Однако, при дальнейшем рассмотрении этой проблемы мы приходим к очевидному барьеру в виде отсутствия необходимых экспериментальных данных по **теплоёмкости** растворяемых веществ. Тем не менее это не означает, что мы можем вовсе пренебречь термодинамическими свойствами молекул в ходе построения датафрейма для обучения.

В качестве альтернативы было принято решение использовать сугубо **динамические свойства** молекул. Т. к. растворимость явным образом коррелирует с полярностью молекулы и растворителя (полярное лучше растворяется в полярном и наоборот), то свойства тем или иным образом «кодирующие» полярность предположительно могут дать хороший результат при обучении. Одним из рассматриваемых свойств, которое было выбрано в качестве кандидата на роль одного из признаков стал **дипольный момент** молекулы. Другими кандидатами также могли стать площадь полярной поверхности, молекулярное силовое поле Мерка, а также количество тех или иных функциональных групп внутри молекулы. Однако ввиду ограниченности вычислительных ресурсов добавление каждого вычисляемого по новому алгоритму признака будет стоить большого времени. Из-за этого было принято решение остановиться на двух свойствах: **дипольном моменте** (как наиболее хорошо и относительно недорого с вычислительной точки зрения характеризующем полярные свойства молекулы), а также **молекулярном весе** вещества (который при прочих равных условиях может быть принят за характеристику размера молекулы).

В случае же рассмотрения молекулы сугубо с точки зрения её внутренней структуры, мы можем опереться на известные алгоритмы хемоинформатики и цифрового представления веществ по их SMILE или SMART-кодировке. Одним из таких алгоритмов канонического представления молекулы является **алгоритм Моргана.** Выбор именно этого алгоритма основан скорее на удобстве использования, нежели чем на его особой эффективности в оцифровке молекул.

Смысл представления молекулы в виде её множества подчастей состоит в том, чтобы оценить вклад каждой из подструктур в значение конкретного свойства (в данном случае – растворимости). Мы предполагаем, что наиболее электроотрицательные атомы (азот, кислород, фтор), а также относительно крупные неполярные части молекул (бензольные кольца, длинные цепи насыщенных связей и т. д.) должны оказывать наибольшее влияние на искомое свойство.

* 1. *Создание датафреймов для обучения моделей.*

Таким образом для обучения моделей и проверки двух концепций были созданы два датафрейма: dipole\_data и bit\_data, скомпонованные на основе дипольного момента и алгоритма Моргана соответственно.

Для создания дипольного датафрейма использовались класс «MoleculeDynamicsDescriptorClass» и функционал библиотеки «openbabel». Для создания побитового датафрейма согласно алгоритму Моргана, были использованы функция «get\_morgan\_bit\_info», класс «CreateMorganBitMatrix» и функционал библиотеки «Rdkit». Ввиду ограниченности вычислительных ресурсов, количество побитовых элементов согласно алгоритму Моргана было принято за 5000 (однако следует участь, что чем меньше это число, тем выше шанс коллизий, когда под одну кодировку в разных молекулах могут попасть разные сайды).

Следует обратить внимание на то, что каждый класс имеет кардинально разную скорость сбора данных. Если для побитового датафрейма это время исчисляется не более чем сотнями секунд, то сбор данных по дипольному моменту занял часы, что связано с «дорогими» с вычислительной точки зрения (по сравнению с остальными) операциями добавления в исходную молекулу атомов водорода и получения непосредственно дипольного момента.

* 1. *Предобработка данных.*

В качестве предобработки данных использован класс «MinMaxScaler» библиотеки «scikit-learn».

Во избежание «утечки данных» в ходе решетчатого подбора параметров для моделей был использован класс «Pipeline». Подбор моделей осуществлялся на отрезках датафреймов размером в 100000 строк.

Подбираемые параметры моделей, а также наилучшие выявленные в ходе подбора результаты представлены в таблицах ниже.

Таблица 1. Параметры подбора Lasso-регрессии.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Использованный датафрейм | Параметры регуляризации | Параметры итерации | Наилучшие подобранные параметры | R2 при перекрестной проверке |
| bit\_data | alpha: [0.001, 0.01, 0.1, 1] | max\_iter: [10, 100, 1000, 10000] | alpha: [0.001]; max\_iter: [100] | 0.47 |
| dipole\_data | alpha: [0.001, 0.01, 0.1, 1] | max\_iter: [10, 100, 1000, 10000] | alpha: [0.001]; max\_iter: [10] | 0.18 |

Лассо-регрессия обладает недостаточно приемлемым качеством для данных наборов данных. Низкие значения R2, вкупе с избыточно высокой степенью регуляризации позволяет сделать вывод о том, что модель L1-регуляризации мало подходит (иначе говоря, количество реально значимых признаков больше, чем отбирает модель).

Таблица 2. Параметры подбора Ridge-регрессии.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Использованный датафрейм | Параметры регуляризации | Параметры итерации | Наилучшие подобранные параметры | R2 при перекрестной проверке |
| bit\_data | alpha: [1, 5, 10, 100] | max\_iter: [100, 1000, 10000] | alpha: [10]; max\_iter: [100] | 0.75 |
| dipole\_data | alpha: [1, 5, 10, 100] | max\_iter: [100, 1000, 10000] | alpha: [1]; max\_iter: [100] | 0.19 |

Гребневая регрессия, использующая L2-регуляризацию показала себя намного лучше на наборе данных, собранном согласно алгоритму Моргана. Средняя регуляризация на этом массиве данных и этой модели говорит об относительной оптимальности подобранного параметра. Тем не менее результаты для набора данных, основанного на дипольном моменте, изменились мало.

Таблица 3. Параметры подбора SVM-регрессии.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Использованный датафрейм | Параметры регуляризации | Параметры итерации | Наилучшие подобранные параметры | R2 при перекрестной проверке |
| bit\_data | С: [0.1, 1, 5, 10] | max\_iter: [100, 1000, 10000] | С: [1]; max\_iter: [10000] | 0.62 |
| dipole\_data | С: [0.1, 1, 5, 10] | max\_iter: [100, 1000, 10000] | С: [0.1]; max\_iter: [10000] | 0.19 |

SVM-модель дала лучший, чем для Лассо-регрессии, но всё ещё малоудовлетворительный результат своей работы.

Таблица 4. Параметры подбора регрессии на основе дерева решений.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Использованный датафрейм | Параметры регуляризации | Параметры критерия дерева решения | Наилучшие подобранные параметры | R2 при перекрестной проверке |
| bit\_data | max\_depth: [2, 5, 10] | criterion: [mse, friedman\_mse, mae] | max\_depth: [10]; criterion: [mae] | 0.22 |
| dipole\_data | max\_depth: [2, 5, 10] | criterion: [mse, friedman\_mse, mae] | max\_depth: [2]; criterion: [mse] | 0.18 |

Регрессия на основе дерева решений показала наихудший результат. И если в случае побитового датафрейма это с большой долей вероятности связано с недостаточной глубиной дерева, то в случае дипольного датафрейма это говорит о том, что признаков попросту недостаточно для того, чтобы построить адекватное дерево решений.

1. **Выводы.**

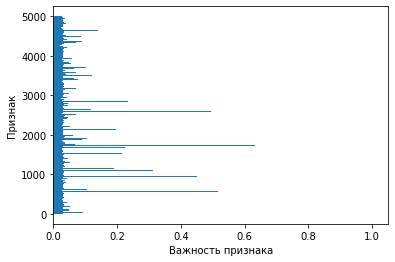
Выбор между данными моделями был основан на компромиссном решении: модели должны быть достаточно быстро обучаемыми (к примеру, линейными), но также хорошо было бы, если бы они могли обогащать изначальный массив данных новыми признаками на основе существующих (ядерный метод опорных векторов) или были легко интерпретируемыми (как деревья решений).

Рассмотрим полученные данные по моделям, построенным на основе дипольного момента молекулы. По большей части, вне зависимости от того, какую именно модель мы используем в ходе обработки данных, мы получаем примерно один и тот же результат: низкий R2 и высокую зарегулированность модели. Можно сделать вывод о том, что количество признаков в данном датафрейме не позволяет построить достаточно точные модели. При наличии необходимых вычислительных ресурсов можно расширить данный датафрейм. Специфика библиотеки «openbabel» позволяет построить не только двумерную структуру молекулы (как было использовано в данном случае), но также и трёхмерную. В этом случае было бы корректно использовать силовое поле молекулы (как характеристику напряженности её связей), площадь полярной поверхности молекулы, трёхмерный дипольный момент, молекулярный вес и количество функциональных заместителей, содержащих в себе наиболее электроотрицательные атомы и / или полярные связи (нитрогруппа, ацильный радикал, гидроксильная группа, галогены и т.д.)

Иначе обстоят дела с моделями, основанными на данных, построенных согласно алгоритму Моргана. Различия результатов, полученных на разных моделях, говорят о том, что непосредственно сам датафрейм имеет достаточно информации для извлечения из него закономерностей. Лучшей моделью в данном случае оказалась гребневая регрессия.

Используя функцию «get\_feature\_importances» мы можем посмотреть важность тех или иных признаков, построенных на основе модели. Зная важность того или иного бита, появится возможность выявить среди них наиболее ценные с информационной точки зрения. Построив простой график на основе важностей мы можем увидеть, что некоторые признаки действительно обладают резко отличающимися от остальных показателями важности:

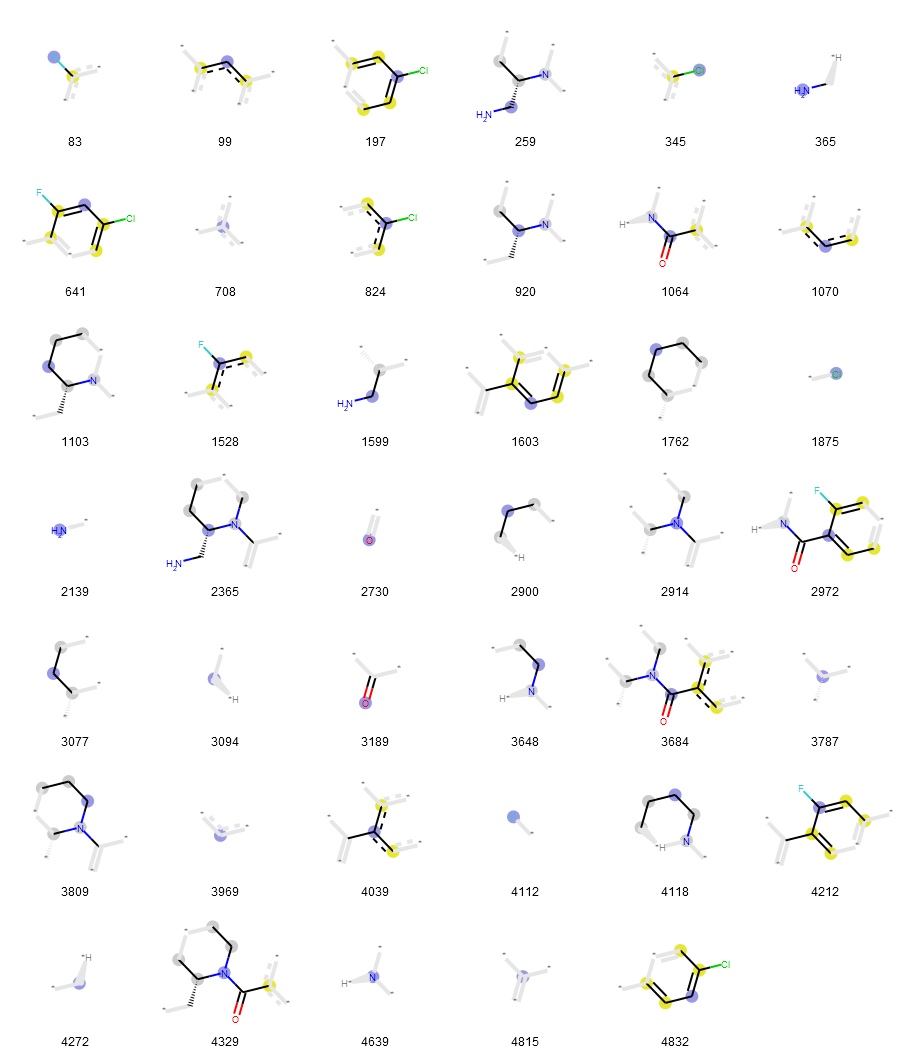
Рисунок 1. Гистограмма важности признаков модели на основе Ridge-регрессии



Проитерировавшись по массиву важности признаков и вычленив из них некоторые наиболее интересные (те, в которых важность оказалась более 0.5), их можно визуализировать, воспользовавшись стандартными средствами генерации изображений библиотеки «Rdkit».

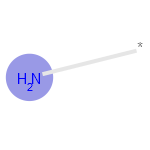
К примеру, молекула с кодировкой «ZINC000044541732» имеет следующий набор бит:

Рисунок 2. Набор бит для молекулы «ZINC000044541732».



В то время как наиболее «значимой» в ней оказалась часть, содержащая аминогруппу.

Рисунок 3. Наиболее «значимая» часть молекулы «ZINC000044541732».



Рассмотрев наиболее значимые биты, можно сделать вывод о том, что части молекул, содержащие в себе азот и аминогруппы, а также бензольные части оказывают наибольшее влияние на растворимость. Наше изначальное предположение частично подтвердилось.

Однако ввиду большого количества изучаемых признаков проверить достоверность этого утверждения не представляется возможным (иначе говоря, не являются ли данные группы лишь кодирующими для менее явного признака или группы признаков, которые собственно и оказывает непосредственное влияние на растворимость?). Ввиду этого, даже несмотря на относительно удовлетворительное качество модели гребневой регресси, построенной на данных собранных посредством алгоритма Моргана, физическая интепретация полученных результатов затруднительна.