Аналитическая записка

Принятые допущения, сознательно опущенные моменты и т.п.

- При решении задачи возникало множество вопросов об особенностях радарных профилей и о предобработке, выполняемой на радаре. На данном этапе я ограничилась доступной мне информацией и своими предположениями, однако в реальной ситуации на этом шаге должна была бы быть подробная консультация с коллегами.

- В процессе предварительного анализа были найдены записи:
  - с отсутствующим радарным профилем (**num_profile** содержит NaN);
  - с одинаковым идентификатором судна, но различными габаритами. Поскольку габариты в датасете явно взяты из внешнего источника, а не рассчитаны на основе данных радара (и в большинстве случаев значения габаритов в записях, соответствующих одному ais_id, одинаковы), я сочла такие записи некорректными;
  - с типом судна *fishery_research vessel*, который встречается в датасете всего один раз, и этого явно не достаточно для идентификации судов данного класса.
  Такие записи были исключены из рассмотрения.

- Также в данных присутствует **класс судна None**. Не стала удалять такие записи. Но если это не один из типов, а неизвестный тип, то такие записи следует удалять.

- При построении модели и формировании признаков я предполагала, что числовые значения, составляющие радарный профиль, представляют из себя амплитуды отраженных сигналов и находятся в некоторой зависимости от расстояний от радара до объекта (и от курсового угла), а также связаны с относительными размерами частей судна, однако эти зависимости неоднозначны, и на результирующие значения могут оказывать влияние множественные отражения сигнала и шумы.

- Поскольку я предположила, что опираться на абсолютные значения амплитуд может быть не совсем корректно (что подтверждается графиками профилей одного и того же корабля под разными углами и на разных расстояниях от радара), то произвела масштабирование радарного профиля на отрезок [0,1] пообъектно. Таким образом, некоторые признаки используют информацию об относительных величинах пиков радарных профилей вместо абсолютных.

- Поскольку радарный профиль изначально зашумлен, скорее всего, необходима фильтрация получаемого сигнала. Возможно, она производится на радаре, а в датасете содержится уже очищенный сигнал. В рамках данного исследования фильтрация сигнала мною не производилась.

- При построении модели была обнаружена сильная зависимость между расстоянием от судна до радара и целевой меткой (***marine_traffic_class***). Я построила модель, которая принимает на вход только два параметра — ***distance_to_radar*** и ***view_angle***. У такой модели получилась точность около 65%. Определение типа судна на основе данных параметров, конечно, смысла не имеет, поскольку данные значения не являются характеристиками исследуемых объектов — судов. Высокую значимость признака

distance_to_radar можно объяснить особенностями собранного набора данных — возможно, суда разных типов чаще всего наблюдались на определенных диапазонах расстояний до радара. Из этого я сделала вывод, что признак ***distance_to_radar*** хоть и может быть полезен при построении классификаторов, однако использовать его напрямую не стоит во избежание потери обобщающей способности моделью. Можно пробовать включать его косвенно — например, при рассчете новых комбинированных признаков, или, может быть, для оценки априорного распределения вероятностей классов (если это нужно, и если появление кораблей разных типов в разных местах акватории не совсем случайно, а может быть каким-то образом обосновано).
По указанным причинам в рамках данного исследования признак ***distance_to_radar*** исключен несмотря на то, что добавление его в модель дает существенную прибавку точности моделей.

Состав и описание решения

eda.ipynb — разведочный анализ данных
features_creation.ipynb — работа с признаками
modeling.ipynb – моделирование

Признаки (подробнее — в коде):

- Исходные радарные профили
- Нормализованные радарные профили
- Значения и индексы пиков, их ранжирование по убыванию
- Соотношения между пиками
- Количество пиков
- Диапазон значений пиков
- Нулевой начальный момент
- Первый начальный момент
- Центр тяжжести и относительное положение центра тяжести
- Второй и третий центральные моменты
- Нормированные центральные моменты
- skewness, kurtosis, медиана

Было подготовлено два датасета — с масштабированием профиля и без.
Не все признаки вошли в модели - некоторые из них ухудшали качество моделей или не давали прироста качества, некоторые были исключены, по причине высокой корреляции с другими использованными признаками.

Для прогнозирования типа судна приведено две модели:
1) RandomForestClassifier показал accuracy 61,73%
   Использован датасет без масштабирования.
2) Нейронная сеть показала accuracy **72,17%**
   Использован датасет без масштабированием.
   Архитектура нейронной сети:
   - полносвязный слой из 218 нейронов с активацией tanh и инициализацией весов glorot-uniform

- слой батч-нормализации
- полносвязный слой из 436 нейронов с активацией ReLU и инициализацией весов Xe
- слой батч-нормализации
- выходной слой с 33 нейронами и активацией софтмакс

Для прогнозирования габаритного класса приведена модель
RandomForestClassifier, которая показала accuracy **77,87%**

Пути для улучшения моделей

- Увеличить количество примеров классов, в которых в текущем датасете мало объектов, путем сбора дополнительных данных, либо путем семплирования (например, с помощью простого oversampling или SMOTE).
  В результате анализа произведенного решения выявились некоторые классы, которые совершенно не распознаются. Это как раз классы с малым количеством примеров. Также следует отдельно их проанализировать и поискать новые признаки, которые позволили бы выделить данные типы судов среди остальных.
- Подобрать архитектуру нейронной сети, которая больше подходит для данной задачи. Например, RBF (радиально-базисные НС)
- Поработать с распределениями признаков. Возможно, некоторые признаки следует прологарифмировать или иным образом обработать.
- Составить новые признаки, в т.ч. использовать расстояние до радара и угол при расчете признаков.
- Использовать фильтрацию для отсечения шумов
- Тщательнее настроить гиперпараметры, использовать кросс-валидацию

  И т.д.