Отчет о решении задачи конкурсу по машинному обучению “Прогноз стартовой обводнённости по ГИС”

1. **Задача конкурса**

Конкурс по машинному обучению проходил с 1 октября 2017 по 1 декабря 2017 года. Участникам на выбор было предложено несколько задач, среди которых была выбрана задача по определению стартовой обводненности скважины по ГИС. Организаторами конкурса задача была сформулирована следующим образом:

*Одной из актуальных задач ГиР является прогноз обводненности по новым скважинам и выделение доли непроизводительной добычи/закачки при эксплуатации скважин, вскрывающих помимо целевого водонасыщенный горизонт.*

*При использовании алгоритма спектрального моделирования геологического строения пласта в качестве промежуточного результата восстанавливается набор каротажных кривых в проектной скважине (любой заданной точке). Эти кривые могут быть использованы для прогноза обводненности на проектируемых к бурению скважинах.*

*Задача – адаптировать самообучающуюся модель на статические данные ГИС (ПС, ГК, ГГКП, АК) и стартовые замеры обводненности для прогноза доли воды в продукции на новых скважинах.*

Организаторами были предоставлены следующие данные:

1. Обучающая выборка – 182 значения обводнённости и более 200 данных РИГИС Шингинского месторождения (для 150 скважин выполнялось соотношение, данные->обводненность), в дальнейшем были предоставлены данные по ГИС для этого же месторождения. Но большинство методик, в связи с поздним выходом данных по ГИС, были обробованы на данных РИГИС.
2. Контрольная выборка – не предоставлялась организаторами. Поэтому было принято решение, при обучении 30% данных оставлять в качестве контрольной выборки для настройки алгоритма.

Кроме самих данных РИГИС и ГИС были предоставлены данные о перфорациях, координатах скважин. Также для получения более точного решения использовались данные о глубинах подошвы и кровли.

1. **Содержание отчета**

Ниже описаны все этапы создания финального решения. Кроме того, в отчете описываются некоторые идеи, которые проверял автор, но они не вошли в финальное решение.

В процессе работы над задачей автор ставил своей целью не просто получить качественное решение, но и проверить как можно больше подходов к решению, причем как подходов к извлечению информации из признаков, так и подходов к алгоритмам классификации.

* 1. **Подход автора к решению задачи**

Задача сведена к признаковой задаче классификации. Сначала данные предобрабатывались: проводилась фильтрация по значениям (из данных удалялись значения nan, также были удалены заведомо неправильные значения (например, для aps значения > 1, для kint, значения < 0, и так далее)). Затем по каждому параметру строилось его признаковое описание. Признаки генерировались на основе разных подходов:

* Аппроксимация данных
* Статистики
* Фурье-анализ(не вошли в финальное решение)

Далее в отчете каждый подход подробно описан и обоснован. Классификация выполнялась регрессионными алгоритмами, которые оценивали по данным каратожа значение обводненности.

Были подробно исследованы возможности следующих алгоритмов:

* Случайный лес
* Бустинг над деревьями
* Различные виды нейросетей
  1. **Реализация подхода**

Технически задача решалась следующим образом. С помощью библиотеки lasio и скрипта на python данные выгружались из las формата в csv (с данными значений параметров по глубине). Затем эти данные загружались в другом скрипте, производилась фильтрация данных, определялись значения кровли и подошвы, и генерировались признаки для задачи обучения, и набору признаков ласа ставились в соответствие значение обводненности (для генерации признаков методом Фурье использовался matlab). Из надора данные выделялись данные для обучения и контроля, в соответсвии 70/30. Данные признаки использовались для настройки регрессоров из библиотеки skit-learn и на нейросетях tensorflow и keras.