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). Данные разделялись для обучения и контроля, в соответствии 70/30. Данные признаки использовались для настройки регрессоров из библиотеки skit-learn и на нейросетях tensorflow и keras.

Общий план работы с данными

1. **Загрузка и анализ данных**

Для первоначального анализа данных, было решено проверить, нет ли явной зависимости значения wc от данных в ласах. Во время анализа данных, было замечено, что лучше анализировать область, лежащую между кровлей и подошвой. Для оценки было решено сравнить, средние значения в этой области для каждого параметра, для установления взаимосвязи между ними (рис. 2). Изначально организаторами конкурса были выданные данные по РИГИС, поэтому первоначальная оценка происходила на этих данных. Значения по параметрам усреднялись в области анализа (кровля-подошва). Из рисунка 2 можно сделать вывод, что по среднему значению, значимой информации выделить не представляется возможным, поэтому было решено, проверить, насколько эффективны различные алгоритмы генерации признаков.

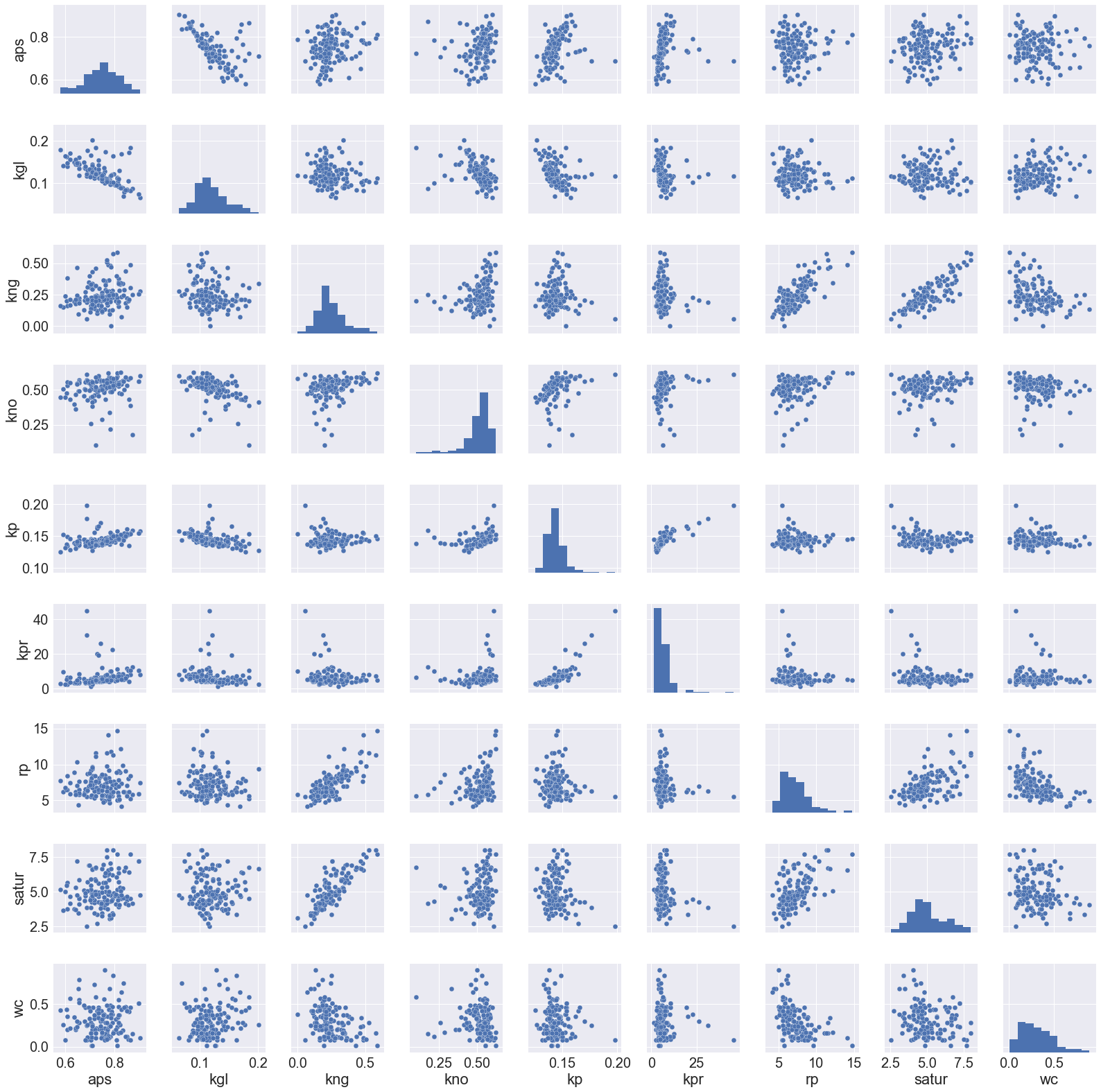


Рис. 2. Корреляция между различными параметрами скважин, для которых были предоставлены значения обводненности, усреднёнными по глубине

1. **Генерация признаков**

Задача классификация признаков сводилась к классической признаковой классификации, когда каждый объект (в данном случая кривая, либо набор кривых) описывался фиксированным набором вещественных признаков. Простейший пример признака – среднее значение на кривой. Однако, признаки должны быть выбраны так, чтобы максимально точно описать сигнал, учесть физику описываемых процессов, также при генерации признаков могут быть использованы различные эвристики.

Для генерации признаков использовалось несколько подходов. Все они описаны ниже. Результатом такой генерации служит некий набор признаков, от десятков до сотен, в зависимости от метода.

Для определения более эффективных методов генерации признаков, проводились эксперименты с настройкой классификаторов на разных признаковых пространствах и оценкой качества классификации на контрольной выборке (30% от общего числа).

1. **Оценка признаков**

Для оценки качества признаков использовался метод MAE (Mean absolute error – средняя абсолютная ошибка):

,

где *wccontrol* – значение обводненности скважины на контрольной выборке, *wcpredict* – к значение обводненности, предсказанное классификатором, *N –* количество скважин в контрольной выборке.

1. **Признаки на основе линейной интерполяции**

Первым был опробован метод на основе кусочно-линейной интерполяциии. Данные по выбранному параметру (кривой), после удаления некорректных значений, рассматривались в промежутке между кровлей и подошвой (в дальнейшем эту операцию будем обозначать - предобработкой), после чего генерировалось значение в k точках, с одинаковым шагом по глубине, между точкам. Далее значения в этих k точках подаются в классификатор.

1. **Статистические признаки**

Сначала каждая кривая предобрабатывалась, затем вычислялись статистические признаки как для самого сигнала, так и для его производных. Генерация на этом этапе проводится в следующем порядке:

1. Предобработка
2. Для сигнала (x1,…,xn), производной (x2 – x1, xn – xn-1) и модуля производной (|x2 – x1|, |xn – xn-1|) вычисляются следующие значения признаков:

* Среднее значение
* Стандартное отклонение
* Доля пересечений с уровнем а (a = 0, a = mean, a = mean + std)
* Значения перцентилей (p10, p50, p90 и т.д.)

1. **Признаки, основанные на разложении Фурье**

При анализе сигналов один из самых успешных методов – использование преобразования Фурье (в нашем случае – одномерного дискретного прямого преобразования Фурье (ДПФ)).

Используя коэффициенты ДПФ добиться хороши результатов в классификации не удалось, поэтому было решено раскладывать не всю кривую, а ее участки. Затем эти разложения усреднить (по логике для получения стабильных признаков). Для этого использовался Matlab, в котором есть функция spectrogram, для разделения сигнала на отрезки и расчета ДПФ на каждом из них. К сожалению, данный метод показал себя хуже первых двух.