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

Авторы: Еникеев М.Р., Фазлытдинов М.Ф.

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 опробованы 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 проверяли авторы, но они не вошли в финальное решение.

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

* 1. **Подходы авторов 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 каротажа значение обводненности.

Были подробно исследованы возможности следующих алгоритмов:

* Случайный лес
* Бустинг над деревьями
* Некоторые 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, масштабированные данные с усреднением значения в этой области для каждого параметра, для установления взаимосвязи между ними (рис. 2). Изначально организаторами конкурса были предоставлены данные по РИГИС, поэтому первоначальная оценка происходила на этих данных. Из рисунка 2 можно сделать вывод, что выделение значимой информации не представляется возможным. Анализ зависимости в логарифмической шкале также не дал результатов. Поэтому было решено, проверить, насколько эффективны различные алгоритмы генерации признаков.

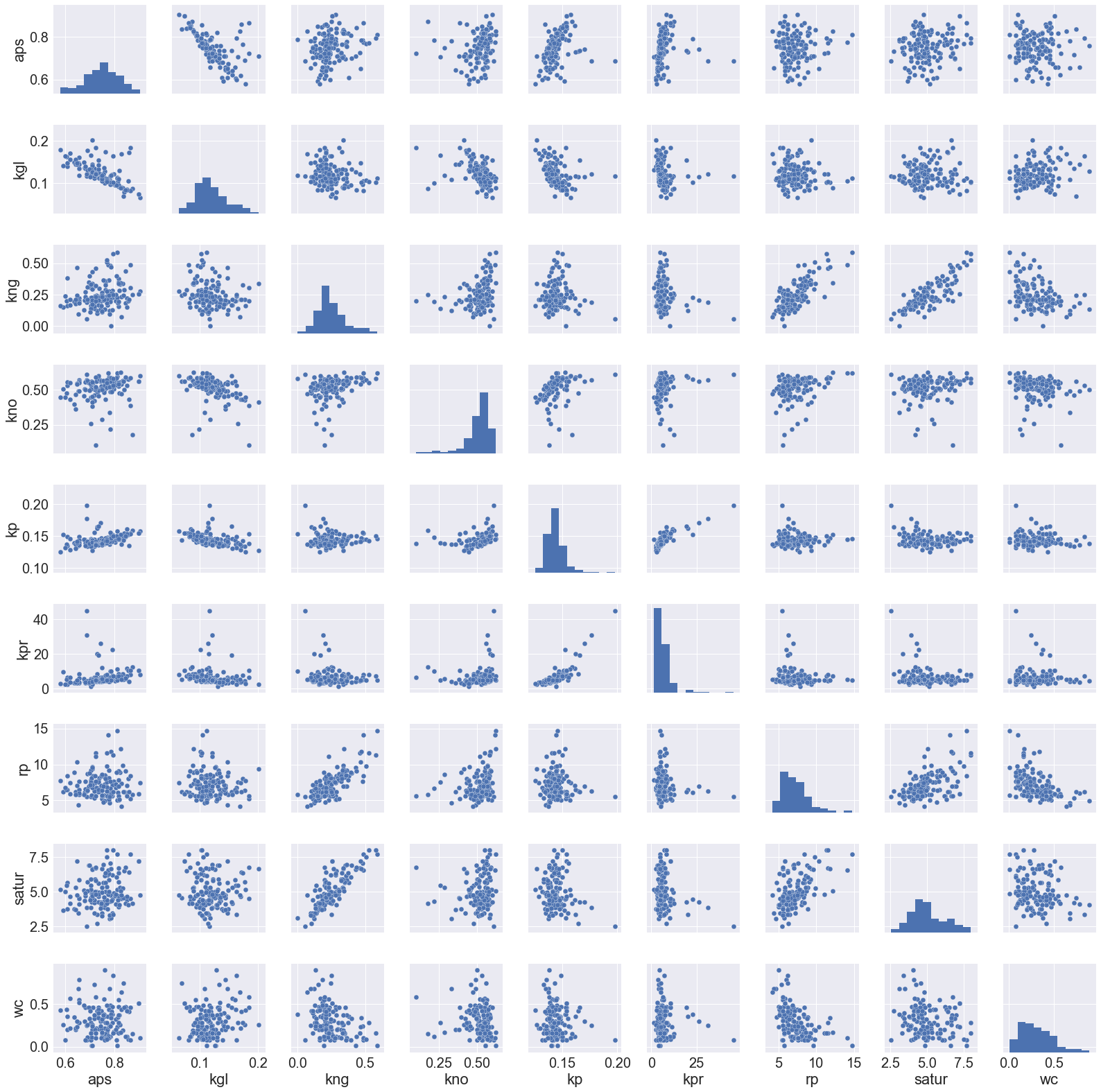


Рис. 2 Корреляция между различными параметрами скважин, для которых были предоставлены значения обводненности, усреднёнными по глубине

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 значений, 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 кусочно-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)

Также был рассмотрен вариант с использованием признаков значений перцентилей (p10, p50, p90 и т.д.), среднее значение и отклонение.

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 хороших 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.

1. **Выбор оптимальных признаков и параметра, по которому проводить обучение**

Для выбора оптимального метода генерации признаком было решено провести классификацию ансамблем деревьев, с поиском оптимальных настроек для метода (изменение максимальной глубины дерева, количества деревьев в ансамбле решений, количество отбираемых признаков).

Для полного анализа оптимальной кривой ГИС для классификации, были проанализированы следующие кривые: 'KINT', 'R05', 'R20', 'R14', 'R10', 'F07', 'F10', 'F14', 'R07', 'F20', 'F05', 'PHIT', 'MRES', 'SG', 'KGL', 'SXWB', 'GZ3', 'NPHI', 'GZ2', 'GZ4', 'GZ1', 'CILD', 'PROX', 'LLD', 'GZ7', 'APS', 'KPS', 'GZ5', так как данные кривые наиболее наполнены данными.

Результаты сравнения работы алгоритмов генерации признаков для большинства кривых представлены на рисунке 3. В результате проведения данного теста, было решено использовать алгоритм, основанный на статистических признаках.

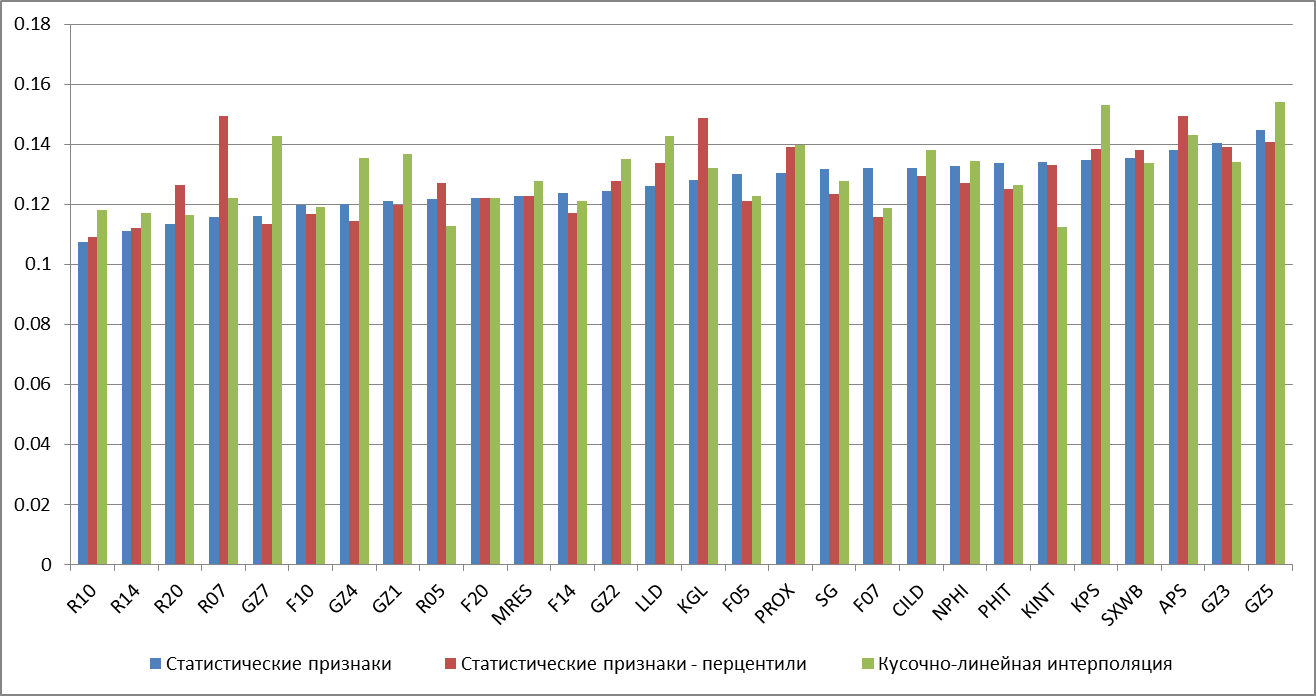


Рис. 3 Сравнение методов генерации признаков для обучения

Среднее значения MAE по разным методам [0.127, 0.128, 0.130], и как мы видим, результат, практически не зависит от метода генерации параметров и от выбора кривой из данных каротажа.

Результат на контрольной выборке (30% от тестовой) по кривой GZ5 представлен на рисунке 4. Признаки рассчитывались статистическим методом. Красным цветом изображены реальные значения обводненности, синим – предсказанные значения.

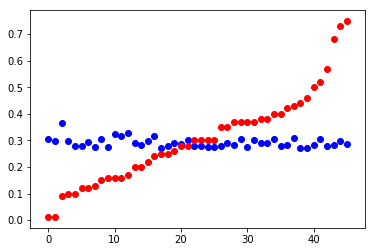


Рис. 4 Работа классификатор по Gz5

Аналогичная картина наблюдается и по другим каротажным кривым. Были проведены тесты по комбинации некоторых каротажных кривых при генерации параметров обучения. Существенного изменения картины предсказания не наблюдалось.

Можно предположить, что полученный результат является следствием неравномерного распределения значений обводненности в обучающей выборке (рис. 5).

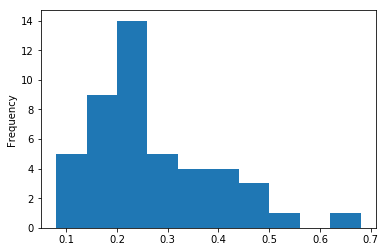


Рис. 5 Гистограммы значений обводненности: для обучающей выборки (слева) и для контрольной выборки (справа)

Было решено проверить, как поведет себя классификатор на более обширной обучающей выборке.

1. **Моделирование данных для увеличения размера обучающей выборки**

Одним из основных требований к задаче для применения методов машинного является большой набор обучающей выборки. Возможные решения проблемы расширения обучающей выборки:

1. Моделирование каротажных данных спектральным методом [автор Исмагилов Нияз Салаватович (если есть статья, тогда лучше ссылаться на нее)] и интерполяция (аппроксимация) карты обводненности.
2. Использование каротажных данных и обводненности нескольким месторождения «схожим» месторождения (например, все месторождения Западной Сибири).

В данной работе было использовано первое решение.

**1.2.1 Моделирование каротажных данных спектральным методом**

Для генерации каротажных данных методом спектрального моделирования необходимы следующие данные:

1. Каротажные данные (Las-файлы, которые были предоставлены организаторами конкурса).
2. Карты кровли и подошвы пласта. Они необходимы для моделирования в пределах пласта.
3. Траектории скважин. Необходимы для сопоставления абсолютных и измеренных глубин.

Для генерации целевых значений обводненности необходима карта обводненности. Она была получена логарифмической аппроксимацией предоставленных данных начальной обводненности.

Описанная схема получения данных каротажных кривых и карты обводненности представлена на рисунке 3.



Рис.3 Схема генерации расширенной обучающей выборки

Результаты каротажных данных полученные с помощью спектрального моделирования представлены на рисунке 4 (синим цветом изображён исходный каротаж, красным – смоделированный) 105 скважина в моделировании не участвовала, и в итоге была корректно предсказана.

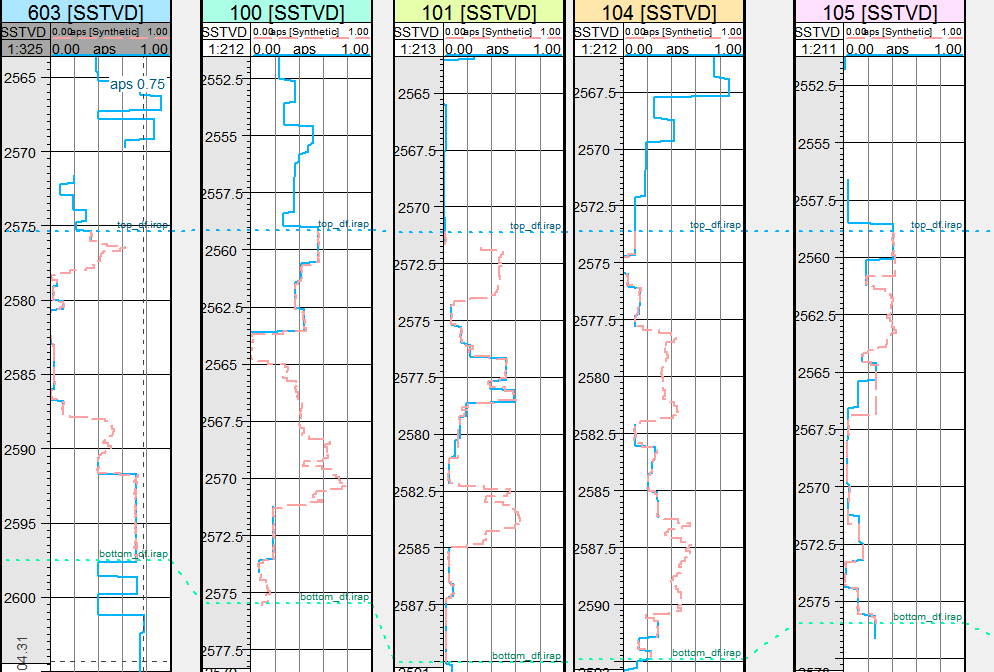


Рис.4 Результаты спектрального моделирования по кривой aps

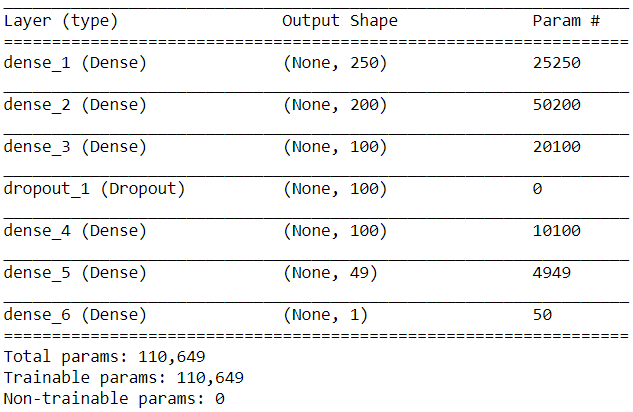
В результате моделирования обучающая выборка была расширена до 5349 скважин. То есть были получены расширенные данные для обучения классификатора.

1. **Обучение на расширенной обучающей выборке**

На данной выборке для генерации признаков использовалась кусочно-линейная интерполяция. В качестве исходных данных рассматривались кривые aps и kgl из РИГИС, так как в рамках конкурсе не было возможности проанализировать и расширить выборку по данным из ГИС.

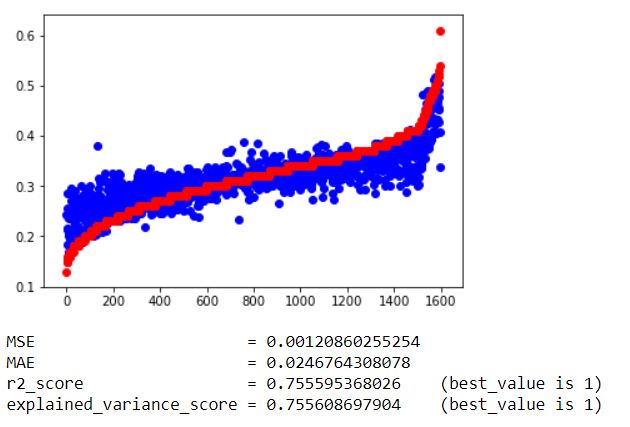
В качестве инструмента классификации были рассмотрены ансамбль деревьев с поиском оптимальных настроек для метода и многослойная нейронная сеть.

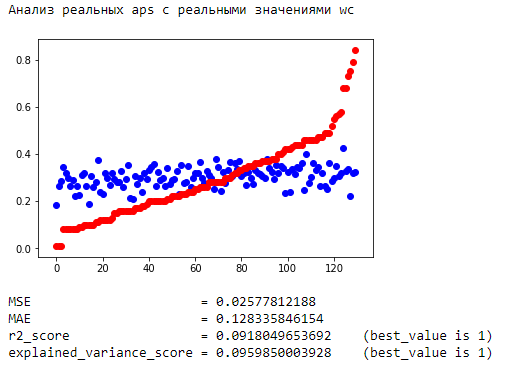
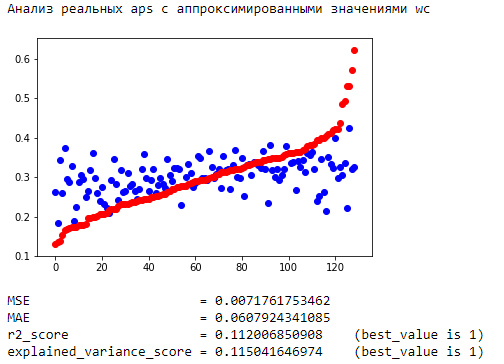
Структура рассмотренной многослойной нейронной сети выглядит следующим образом



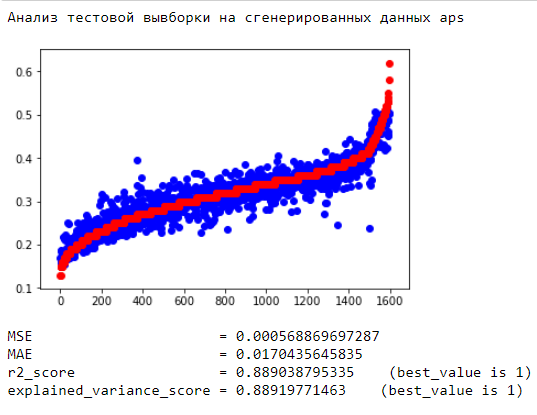
Обученная модель проверялась на трех типах данных: контрольная выборка (выделенная из расширенного обучающего набора), значения wc на апроксимированной карте (для исходных скважин) и реальные значения wc (для исходных скважин).

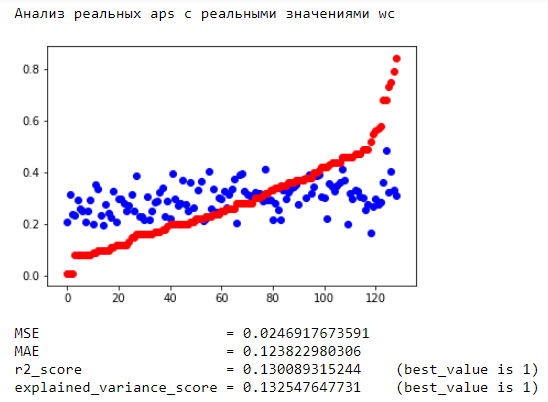
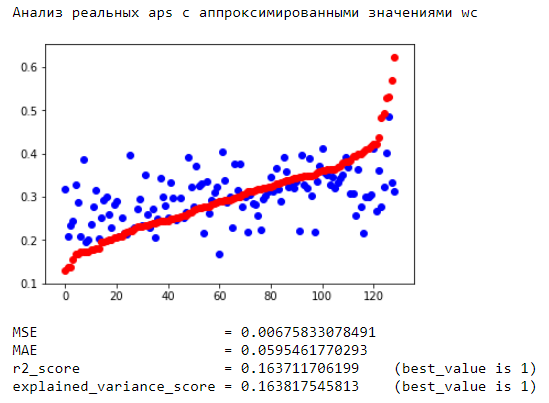
Результаты классификации на данных aps (ансамбль деревьев):



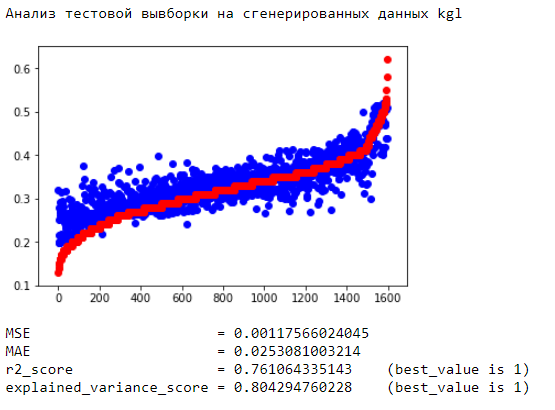


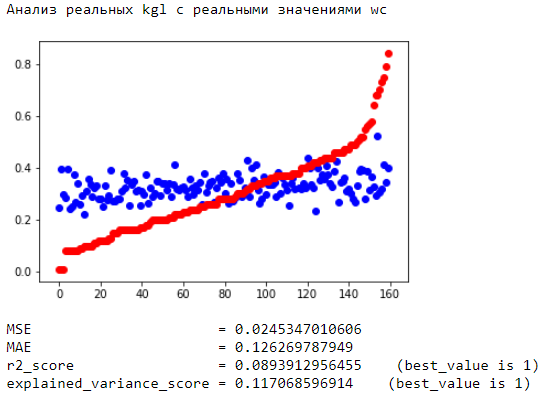
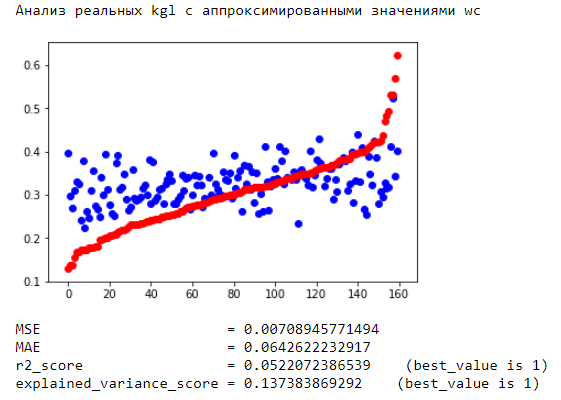
Результаты классификации на данных aps (нейросеть):



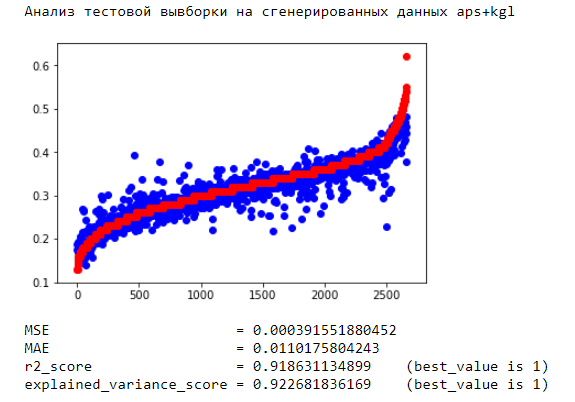


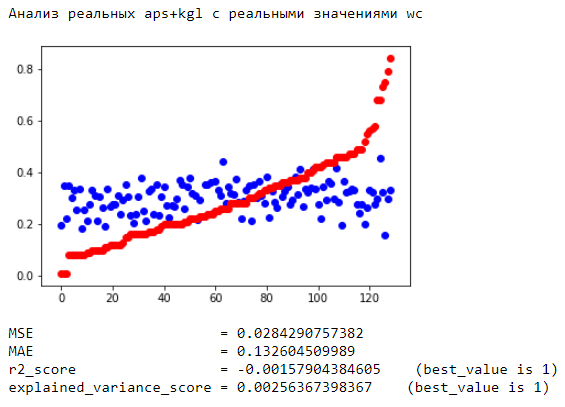
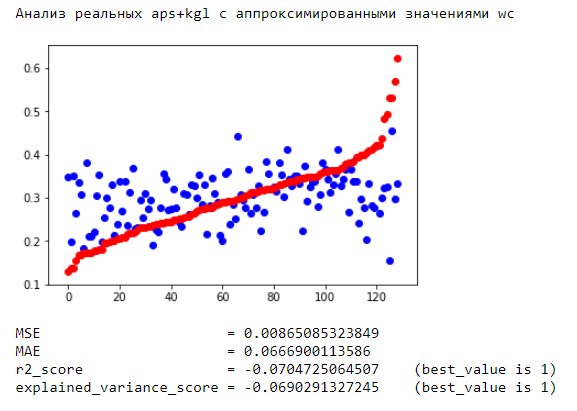
Результаты классификации на данных kgl (нейросеть):





Результаты классификации на комбинации данныхaps и kgl (нейросеть):





Как мы можем видеть, все модели показали результат одного порядка. Высокую точность на контрольной выборке, и гораздо худший прогноз на реальных скважинах.