Санкт-Петербургский Государственный Университет
Математическое обеспечение и администрирование
информационных систем
Информационные системы и базы данных

Орнатская Анна Александровна

Выпускная квалификационная работа бакалавра

Анализ применения методов машинного обучения к задаче каротажа

Научный руководитель :
к. ф.-м. н., доцент Графеева Н.Г.

Рецензент:
заместитель начальника отдела УСИТ СПбГУ
Григорьева Л.И.

Санкт - Петербург
2018

Saint Petersburg State University
Software and Administration of Information Systems
Information Systems and Data Bases

Ornatskaia Anna

Graduation Project

Analysis of the application of machine learning methods to the logging problem

Scientific supervisor:
PhD, associate professor Natalia Grafeeva

Reviewer:
Deputy head of IT Department Ludmila Grigorieva

Saint Petersburg
2018

Оглавление

Введение 4

Терминология 5

1. Постановка задачи 6

2. Обзор существующих результатов 7

3. Исходные данные 10

4. Предварительная обработка данных 14

4.1. Устранение шумов в измерениях каротажей 14

4.2. Нормировка 15

4.3. Восстановление неизмеренных значений 17

5. Применение методов машинного обучения к задаче каротажа 20

5.1. Логистическая регрессия 21

5.2. Метод опорных векторов 23

5.3 Наивный Байес 25

5.4. Деревья решений 25

5.5. Метод ближайших соседей 26

5.6. Результаты 28

6. Постобработка данных 32

Заключение 34

Список литературы 36

Приложение 1 38

Приложение 2 39

Введение

Постоянный рост объемов разработки новых месторождений ископаемых способствуют поиску новых способов выполнения этого процесса, которые смогли бы повысить его эффективность. Экономические и производственные показатели добычи ископаемых напрямую зависят от точности и быстроты интерпретации геологоразведочных данных.

Классическим подходом к определению горных пород в новом месторождении является получение керна (цилиндрический монолит горной породы, получаемый при бурении поисковых и разведочных скважин). Но, к сожалению, такой способ не всегда технически возможен, а также невыгоден экономически [2]. Для решения этих проблем применяется альтернативный метод получения геологической информации — использование геофизических исследований. При его использовании пропадает необходимость отбора керна. Основой метода является каротаж (геофизические исследования скважин, выполняемые с целью изучения геологических разрезов и выявления полезных ископаемых). С его помощью, путем измерения широкого спектра свойств горных пород, можно получить их подробное описание на различных глубинах. В настоящее время бурение любой скважины обязательно сопровождается такими исследованиями.

Зависимости каротажных измерений от соответствующих им пород часто имеют сложный нелинейный характер. Это позволяет сделать предположение об эффективности применения для решения такой задач методов машинного обучения.

В последние годы повсеместно внедряется современная геофизическая аппаратура, позволяющая оперативно производить комплексную обработку и интерпретацию геофизической информации с помощью компьютеров. Задача автоматической интерпретации является крайне актуальной с учетом большого числа пробуриваемых скважин и требований по оперативной обработке данных.

Терминология

Фация - порода, возникающая в определённой обстановке.

Каротаж - геофизические исследования скважин, выполняемые с целью изучения геологических разрезов и выявления полезных ископаемых.

Каротажная диаграмма - диаграмма результатов геофизических исследований в скважине, представляющая собой кривые изменения физ. параметров (или показаний скважинных приборов) по разрезу скважины (Рис. 1).

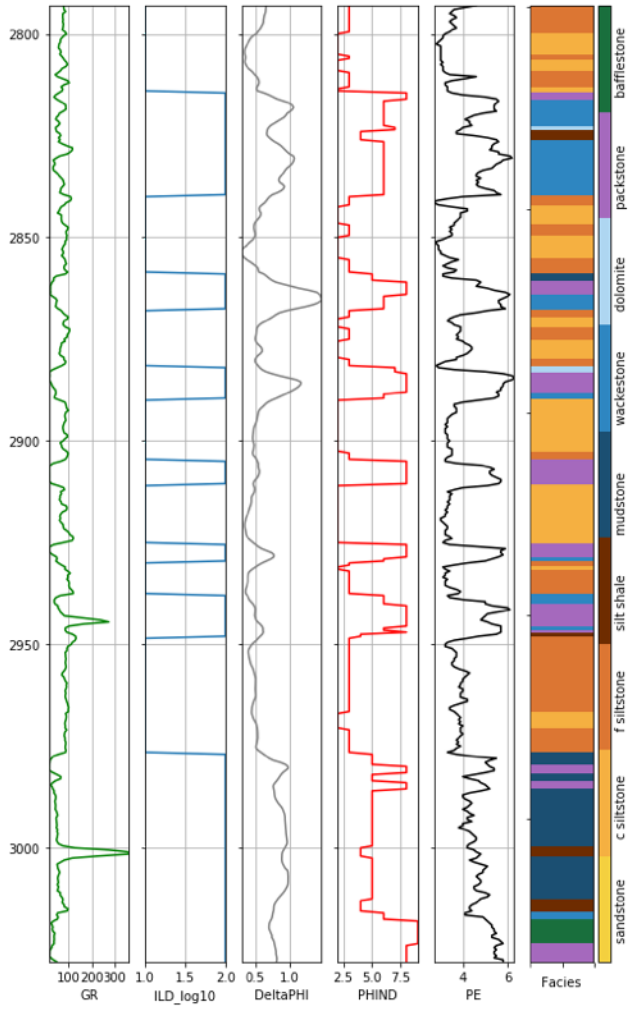


Рисунок 1. Каротажная диаграмма для скважины

Каротажная диаграмма показывает измерение геофизических свойств (GR, ILD_log10, DeltaPHI, PHIND, PE) фаций вдоль ствола скважины(вертикальная ось - глубина). В правом столбце указаны фации, залегающие на той или иной глубине.

1. Постановка задачи

Целью данной работы является создание эффективного способа определения фаций на основании данных геофизических исследований (данные каротажа).

Для этого необходимо выполнить следующие задачи:

1. Найти подходящий набор исходных данных (каротажные измерения).
2. Найти наиболее подходящие способы предобработки данных каротажа (с учетом особенностей предметной области).
3. Применить алгоритмы машинного обучения к обработанным каротажным данным.
4. Обработать результаты, полученные в пункте 3 (с учетом особенностей предметной области).
5. Проанализировать полученные результаты.

2. Обзор существующих результатов

Проблеме автоматического определения фаций на основе каротажных данных было посвящено несколько работ, освещающих различные методы решения вопроса.

В работе [12] авторы уделили особое внимание выделению нефтяного песка (среди сланца, соляного и нефтяного песков) на основе данных каротажа скважины и сейсмических данных с использованием алгоритмов машинного обучения. Для поиска наилучшего решения были рассмотрены 2 линейных классификатора (Softmax Regression and Gaussian Discriminant Analysis) и 2 нелинейных классификатора (Random Forest and Support Vector Machine). Все модели хорошо зарекомендовали себя на различном песке из сланца, хотя дифференциация нефтяного песка из соляного песка оказалась немного менее оптимальной. Как и ожидалось нелинейные модели превзошли линейные модели, о чем свидетельствует их более низкая ошибка в обучении и тестировании. Среди всех четырех алгоритмов метод опорных векторов достиг лучших результатов классификации (ошибка составила 0.051 на обучающей выборке и 0.101 на тестовой), особенно в отношении небольшого числа false-positive прогнозов нефтяного песка.

В работе [13] рассматривается эффективность различных функций ядра SVM (Support Vector Machine) в предсказании состава пород. Наилучшим полученным результатом является ошибка классификации в 14.21%, полученная на тренировочной выборке, составляющей 55.50% от всех данных. Этот результат был достигнут при использовании полиномиального ядра. Однако породы, имеющие схожие петрофизические свойства не были классифицированы верно (например, ангидрит в большинстве случаев был предсказан как доломит).

В работе [5] предложен способ решения задачи классификации, основанный на градиентном бустинге, представляющим собой совокупность деревьев решений. Для решения проблемы невозможности линейного разделения на кластеры и упрощения работы классификатора, авторами было

предложено: использовать некоторые дополнительные функции (с помощью добавлений нелинейностей) для лучшего разделения на классы в расширенном пространстве объектов, считать, что фации в соседних слоях сильно коррелированы, а также после обработки данных устранять ложные изолированные значения. Результаты тестирования показали, что описанные выше дополнения действительно повышают производительность, точность классификатора: часто встречающиеся в обучающей выборке фации определились верно в среднем в 75% случаев (однако, редко встречающиеся фации классифицируются ошибочно), F-мера составила 0.61 с добавлением дополнений и 0.55 без них.

В работе [3] кроме применения алгоритмов машинного обучения для определения фаций были рассмотрены некоторые способы устранения шумов и других неинформативных составляющих, вызванных спецификой изучаемой области. Автор рассмотрел влияние двух дискретных преобразований: анализа Фурье и вейвлет-преобразования, способствующих улучшению определения фаций. Из-за маленькой вероятности присутствия высокочастотных шумов в каротажных данных применение преобразования Фурье может привести к потере информативности сигнала(перестанут быть заметны флуктуации пород). Однако было замечено, что данный способ сглаживания может положительно повлиять на удаление низкочастотных шумов, которые могут возникнуть с изменением амплитуды сигнала, связанной с закислением пород. Для сглаживания нестационарных шумов было рассмотрено вейвлет-преобразование. В работе был применен вейвлет Добеши. Данный метод сглаживания либо никак не отразился на улучшении качества классификации, либо только ухудшил результат. Однако было замечено, что вейвлет-преобразование подходит для упрощения распознания экстремумов, информация о которых необходимы при интерпретации данных каротажа. Таким образом, рассмотренные преобразования не принесли значительных улучшений или вовсе привели к ухудшению результатов.

Непосредственно определение фаций было произведено с помощью алгоритмов: LDAC (Linear Discriminant Analysis Classifier), SVM (Support

Vector Machine), DLDA (Diagonal Linear Discriminant Analysis), k-NN (алгоритм k ближайших соседей), искусственные нейронные сети (ИНС). Лучшую точность классификации, равную 59%, удалось достигнуть при использовании k-NN (50 соседних точек) и ИНС. ИНС прямого распространения обеспечила лучшую обучаемость (лучшее качество кривой обучения).

3. Исходные данные

Для применения методов машинного обучения был использован набор открытых данных каротажных измерений, представленный на сайте Университета Канзаса (http://www.people.ku.edu/~gbohling/EECS833/facies_vectors.csv). Данные получены из газового месторождения в Юго-Западном Канзасе (10 скважин, 4149 измерений, взятых с интервалом в полфута).

В таблице 1 описаны представленные в исходных данных фации.

Таблица 1. Справочник фаций

Номер фации	Обозначение фации	Название фации
1	SS	неморской песчаник
2	CSiS	неморской крупный алевролит
3	FSiS	неморской мелкий алевролит
4	SiSh	морские алевролит и сланец
5	MS	аргиллит
6	WS	ваккит
7	D	доломит
8	PS	пакстоун-грейнстоун
9	BS	phylloid-algal bafflestone

Некоторые фации обладают схожими характеристикам и нечеткими границами («смежные» фации). Это является причиной постепенного смешивания таких фаций. В таблице 2 для каждой фации представлены «смежные» ей.

Таблица 2. Смежность фаций

Фация	«Смежные» фации
SS	CSiS
CSiS	CSiS, FSiS

FSiS	CSiS
SiSh	MS
MS	SiSh, WS
WS	MS, D
D	WS, PS
PS	WS, D, BS
BS	D, PS

Кроме того, исходные данные включают в себя результаты каротажных исследований, описанных в таблице 3.

Таблица 3. Справочник измерений

Обозначение измерения	Название измерения	Краткое описание
Facies	Номер фации	Определяет фацию (согласно справочнику фаций)
Well Name	Название скважины	Название скважины, на которой были проведены исследования
Depth	Глубина	Глубина(в футах), на которой был измерен признак
GR	Гамма-излучение	Гамма-каротаж: измеряет естественную радиоактивность образования
ILD_log10	Удельное сопротивление	Каротаж сопротивления: измеряет подповерхностную способность препятствовать течению электрического тока
PE	Фотоэффект	Гамма-гамма-каротаж: измеряет излучение электронов фаций, освещенных световыми лучами
DeltaPHI	Разность пористости нейтронной плотности	Нейтронный гамма-каротаж: коррелирующие с плотностью фаций измерения
PHIND	Средняя пористость нейтронной плотности	
NM_M	Индикатор «неморская-морская фация»	Показывает морская или неморская фация

RELPOS	Относительное положение	Целочисленный индекс каждой глубины, на которой было проведено измерение (начиная с 1 для нижнего слоя, увеличивается с глубиной)

На рисунке 2 представлен фрагмент исходных данных.

	Facies	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS
0	3	SHRIMPLIN	2793.0	77.450	0.664	9.900	11.915	4.600	1	1.000
1	3	SHRIMPLIN	2793.5	78.260	0.661	14.200	12.565	4.100	1	0.979
2	3	SHRIMPLIN	2794.0	79.050	0.658	14.800	13.050	3.600	1	0.957
3	3	SHRIMPLIN	2794.5	86.100	0.655	13.900	13.115	3.500	1	0.936
4	3	SHRIMPLIN	2795.0	74.580	0.647	13.500	13.300	3.400	1	0.915
...	...	...	...	...	...	...	...	...	...	...
4144	5	CHURCHMAN BIBLE	3120.5	46.719	0.947	1.828	7.254	3.617	2	0.685
4145	5	CHURCHMAN BIBLE	3121.0	44.563	0.953	2.241	8.013	3.344	2	0.677
4146	5	CHURCHMAN BIBLE	3121.5	49.719	0.964	2.925	8.013	3.190	2	0.669
4147	5	CHURCHMAN BIBLE	3122.0	51.469	0.965	3.083	7.708	3.152	2	0.661
4148	5	CHURCHMAN BIBLE	3122.5	50.031	0.970	2.609	6.668	3.295	2	0.653

Рисунок 2. Фрагмент исходных данных

В исходных данных отсутствуют единичные пропуски, но в трех из десяти скважин («ALEXANDER D», «KIMZEY A», «Recruit F9») не были измерены значения фотоэффекта (PE). Эти пропуски составляют 23-24% измерений (от 415 до 466 неизмеренных значений подряд), что делает невозможным заполнение этих значений соседними. Таким образом, для решения основной задачи определения фаций необходимо решить вспомогательную подзадачу восстановления неизмеренных данных. Для решения данной проблемы применялись методы машинного обучения.

Определение фаций также происходило с использованием методов машинного обучения.

Для применения методов машинного обучения к вышеописанным задачам на основе исходных данных были созданы следующие тренировочные и тестовые наборы:

1) для задачи восстановления неизмеренных значений фотоэффекта: в

качестве тренировочного и тестового наборов были взяты данные скважин, имеющих значение PE («SHRIMPLIN», «SHANKLE», «LUKE G U», «CROSS H CATTLE», «NOLAN», «NEWBY», «CHURCHMAN BIBLE»), в соотношении 80% к 20% соответственно. В наборах использовались следующие признаки: GR, ILD_log10, DeltaPHI, PHIND, NM_M и RELPOS.

2) для задачи определения фаций по каротажным данным с помощью алгоритмов машинного обучения: в качестве тестового набора были выбраны данные скважины «KIMZEY A», так как она содержит в себе все представленные виды фаций (благодаря этому можно оценить распознаваемость каждого вида фаций), а также на ней не производились измерения фотоэффекта (наличие восстановленных значений сделает задачу более общей). Данные остальных 9 скважин стали обучающей выборкой. На графиках 1 и 2 показаны распределение фаций в тренировочном и тестовом наборе задачи определения фаций соответственно.

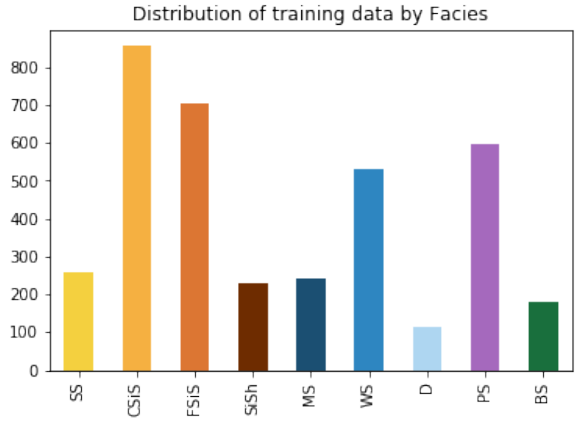


График 1. Распределение фаций в обучающей выборке.

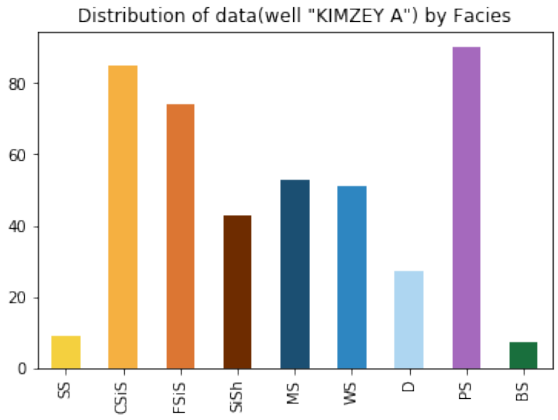


График 2. Распределение фаций в тестовой выборке (скважина «KIMZEY A»).

4. Предварительная обработка данных

Для создания продуктивных моделей прогнозирования необходимо быть уверенным в качестве исходных данных. В реальной жизни они могут содержать в себе ошибки, пропуски, неточности, способные испортить результаты в дальнейшем.

Для повышения качества данных и, как следствие, эффективности модели и улучшения результатов обработки каротажа были применены следующие методы предварительной обработки данных:

1) Устранение шумов в каротажных данных.
2) Нормировка
3) Восстановление неизмеренных значений в данных.

4.1. Устранение шумов в измерениях каротажей

Результаты практически всех методов геофизических исследований в большинстве случаев отличаются высоким уровнем естественных шумов и статистических флюктуаций измеряемых величин. Это связано с их физической природой, неоднородностями природных и геологических сред, закислением скважин и неточностями приборов для регистрации данных [1], [3].

Шумы несут в себе дезинформацию, которая может негативно отразиться на результатах дальнейших исследований. Для устранения шумов был применен метод эмпирической модовой декомпозиции (EMD), описанный в [11]. В отличие от часто используемых для устранения шумов и неинформативных составляющих в геофизических данных методов, таких как преобразования Фурье и вейвлет-преобразования, в процессе эмпирической модовой декомпозиции производится разложение на эмпирические моды, которые не заданы аналитически и определяются исключительно самой анализируемой последовательностью. При этом базисные функции преобразования формируются адаптивно, непосредственно из входных данных.

В основе метода декомпозиции лежит предположение, что исследуемые данные состоят из различных колебательных процессов, при этом в любой момент времени сигнал может содержать множество различных колебательных процессов, нанесенных друг на друга. В общем случае, алгоритм состоит из последовательных операций по выделению модовых функций из сигнала, начиная с высокочастотных. В итоге получается, что исходный сигнал раскладывается по адаптивному базису, полученному из анализируемых данных [4].

Метод эмпирической модовой декомпозиции был применен ко всем каротажным измерениям. На графике 3 представлен пример применения метода эмпирической модовой декомпозиции к каротажным данным.

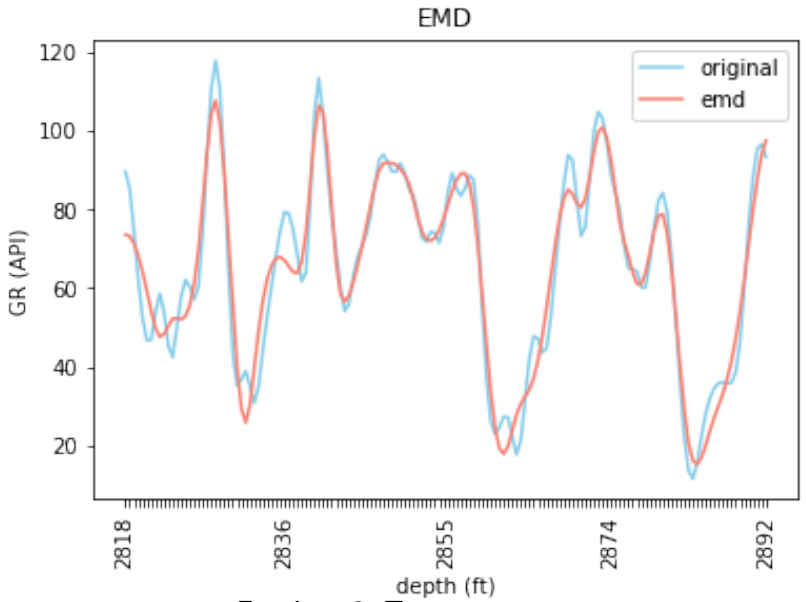


График 3. Пример применение EMD к сигналу гамма-излучения.

4.2. Нормировка

Различные в физическом смысле данные часто различаются и по абсолютным величинам (график 4). Так например, значения гамма-излучений измеряются в API (единица скорости счёта при гамма-каротаже Американского нефтяного института), а значения фотоэффекта в eV (электронвольты), что делает некорректным сравнение их абсолютных величин. Методы машинного обучения чувствительны к масштабированию данных. Нормировка нивелирует большой разброс исходных данных.

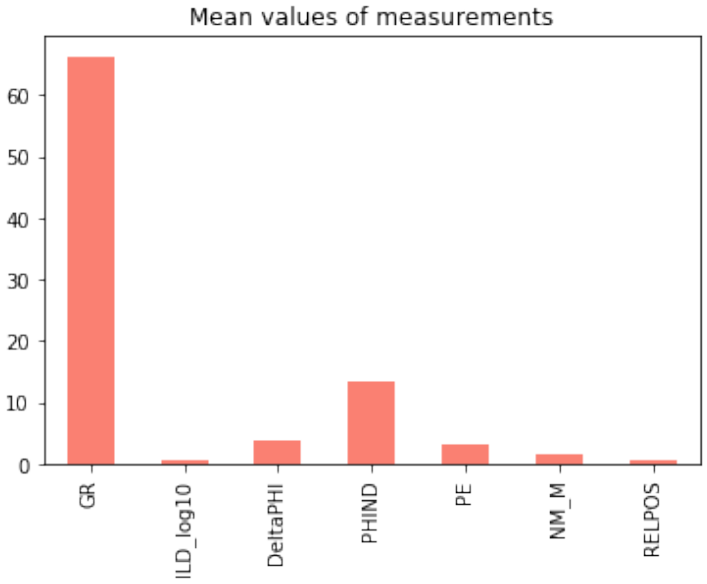


График 4. Средние значения измерений.

В целях улучшения качества моделей машинного обучения были рассмотрены следующие способы нормировки данных:

1. Линейная нормализация

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

2. Стандартизация

$$x_i = \frac{x_i - M(x_i)}{q(x_i)}$$

Где M - математическое ожидание, q - среднеквадратическое отклонение

3. Нелинейная нормировка(сигмоида).

$$x_i = \frac{1}{e^{-a(x_i - x_c)} + 1}$$

Где a - степень нелинейности изменения переменной

$$x_c = (x_{min} + x_{max})/2$$

4. Нелинейная нормировка(гиперболический тангенс).

$$x_i = \frac{e^{a(x_i - x_c)} - 1}{e^{a(x_i - x_c)} + 1}$$

Представленные выше способы нормировки были применены к исходным данным. Для каждой задачи были подобраны наиболее подходящие из них.

4.3. Восстановление неизмеренных значений

В исходном наборе данных значение фотоэффекта не известно для 23-24% измерений (от 415 до 466 неизмеренных значений подряд). Эти пропуски могут повлечь за собой ухудшение дальнейшего результата работы. Для устранения отсутствующих значений были применены методы машинного обучения. Для построения регрессионной модели были опробованы следующие методы:

1. Метод наименьших квадратов.
2. Случайный лес.
3. Логистическа регрессия.
4. Метод ближайших соседей.
5. Опорный вектор регрессии.

Вышеперечисленные методы были применены к тренировочному набору для задачи восстановления неизмеренных значений. Более подробное их описание представлено в [7], [9], [10], [15], [16]. Для определения наиболее подходящего метода были использован коэффициент детерминации (формула 1), вычисленный на тестовой выборке задачи заполнения неизмеренных значений.

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y}_i)^2} \qquad (1)$$

Где y- реальные значение, $\hat{y}$ - предсказанные значения, $\bar{y}$ - среднее по всем реальным значениям y.

Параметр для нелинейных нормировок был подобран экспериментально.

Таблица 4. R2 методов машинного обучения на различно нормированных данных

	Метод наименьших квадратов	Случайный лес (200 деревьев)	Метод ближайших соседей (4 соседа)	Опорный вектор регрессии (линейное ядро)	Логистичес - кая регрессия
Ненормированные значения	0,5	0,74	0,55	0,42	0,27
Линейная нормализация	0,5	0,67	0,52	-0.05	-0,2
Стандартизация	0,5	**0,76**	0,75	0,45	0,28
Нелинейная нормировка (сигмоида, a = 1,2)	0,45	0,63	0,59	0,02	0,23
Нелинейная нормировка (гиперболический тангенс, a=1,2)	0,45	0,63	0,59	0,15	0,25

На графике 5 представлен пример восстановленных данных фотоэффекта.

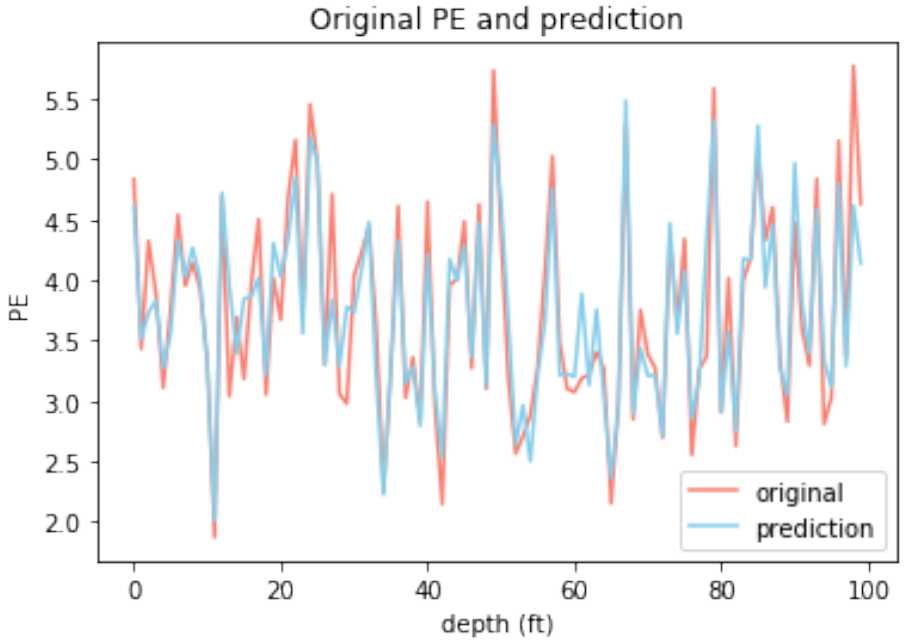


График 5. Пример исходных и восстановленных значений фотоэффекта.

Лучший результат показал стандартизированный набор исходных данных со значениями фотоэффекта, восстановленными с помощью метода случайного леса. Этот способ заполнения неизмеренных значений был использован далее для решения задачи определения фаций по её

геофизическим признакам (при восстановлении значений не производилась нормировка значений фотоэффекта из тренировочного набора, так как необходимо было получить приближенные к реальным значения PE, чтобы в дальнейшем была возможность подобрать наиболее эффективный способ нормировки для задачи определения фаций).

5. Применение методов машинного обучения к задаче каротажа

Для повышения эффективности разработки скважин возникла идея определять находящиеся в них фации только на основании результатов каротажей, произведенных на этой скважине. Для решения этой задачи классификации были рассмотрены несколько методов машинного обучения.

Для построения модели были использованы следующие методы:

1) Логистическа регрессия.

2) Метод ближайших соседей.

3) Метод опорных векторов.

4) Наивный Байес.

5) Деревья решений.

Вышеописанные методы были применены к очищенным от шума и неинформативных составляющих исходным данным с восстановленными неизмеренными значениями.

Результаты были оценены с помощью следующих метрик:

1. Accuracy*

$$Accuracy = \frac{P}{N}$$

Где P - количество правильно классифицированных значений, N - размер обучающего набора

* Далее в работе для замены слова accuracy будет использоваться слово точность.

2. Precision

$$Precision = \frac{TP}{TP + FP}$$

3. Recall

$$Recall = \frac{TP}{TP + FN}$$

4. F

$$F = 2\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

Где P - количество истина-положительных решений, FP - количество ложно-положительных решений, FN - количество ложно-отрицательных решений.

Прежде чем начать определять значения фаций, необходимо нормировать исходные данные. Для определения наиболее подходящего способа нормировки тренировочный набор задачи определения фации был разбит на тренировочную и тестовую(калибровочную) подвыборки в соотношении 80% к 20% соответственно. В таблице 5 показаны результаты применения различных способов нормировок для калибровочной подвыборки (достигаемое на калибровочной выборке accuracy). Параметры для нелинейных нормировок был подобран экспериментально.

Таблица 5. Accuracy

	Метод ближайших соседей	Логистическая регрессия	Метод опорных векторов	Наивный Байес	Деревья решений
Ненормированные значения	0,37	0,59	0,57	0,29	0,65
линейная нормализация	0,4	0,42	0,32	0,42	0,58
стандартизация	**0,56**	**0,6**	**0,66**	**0,48**	**0,66**
нелинейная нормировка (сигмоида)	0,47	0,47	0,47	0,12	0,62
нелинейная нормировка (гиперболический тангенс)	0,47	0,5	0,48	0,12	0,62

Для применения алгоритмов к исходному набору данных была применена стандартизация. Далее рассмотрим подробнее алгоритмы машинного обучения. Для алгоритмов были подобраны параметры. Лучшие из них были определены с помощью метрики accuracy.

5.1. Логистическая регрессия

Идея логистической регрессии заключается в оценке вероятности принадлежности объектов к тому или иному классу по значениям множества признаков. Для этого вводится зависимая переменная y, принимающая одно

из двух значений: 0 (если событие не произошло) и 1 (если событие произошло), и множество независимых переменных (признаков) $x_1, x_2, \ldots, x_n$. На основе значений последних вычисляется вероятность принятия того или иного значения зависимой переменной. Вероятность наступления y: $P(y \mid x) = f(z)^y(1-f(z))^{1-y}$, где $y \in \{0,1\}$, $z = \theta^T x = \theta_1 x_1 + \ldots + \theta_n x_n$, где x и θ - вектор столбцы $x_1, \ldots, x_n$ и коэффициентов регрессии (вещественные числа $\theta_1, \ldots, \theta_n$) соответственно, f(z) - логистическая функция (2).

$$f(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Более подробно метод описан в [15]. Для многоклассовой классификации применяется подход one-vs-rest.

Эффективность логистической регрессии зависит от параметра регуляризации C и вида регуляризации penalty ($C \in (0,11]$, penalty $\in \{l1,l2\}$). Комбинация параметров была подобрана с помощью кросс-проверки, и была выбрана та, которая проявила на ней себя лучше других. Обучения проводилось на тренировочной подвыборке задачи, accuracy измерялась на калибровочной подвыборке. Зависимость точности классификации от параметров представлена на графике 6.

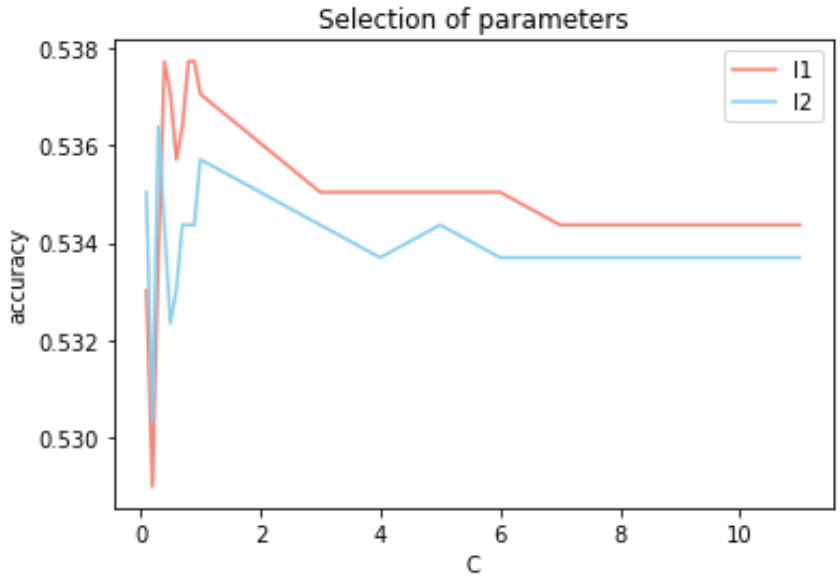


График 6. Зависимость точности от параметров логистической регрессии.

Самую высокую точность (53.7% на калибровочной выборке) модель достигла при использовании параметра регуляции 0.8 и penalty l1. Финальная модель, используемая для установления фаций, обучается затем на тренировочной выборке задачи определения фаций с использованием выбранных параметров.

5.2. Метод опорных векторов

Суть метода опорных векторов заключается в построении гиперплоскости или набора гиперплоскостей в многомерном или бесконечномерном пространстве. Интуитивно, хорошее разделение достигается гиперплоскостью, которая имеет наибольшее расстояние до ближайших точек тренировочных данных любого класса (так называемый отступ). В целом чем больше отступ, тем ниже ошибка классификатора. Более подробно метод описан в [17]. Для многоклассовой классификации применяется подход one-vs-one.

Эффективность метода опорных векторов зависит от выбора ядра (kernel), коэффициента ядра (ширины) (gamma) и коэффициента регуляризации (C). Параметр C определяет компромисс между количеством ошибок на обучающей выборке и простотой линейного решающего правила. С уменьшением параметра C, поверхность решения становиться более гладкой, в то время как увеличение C нацелено на правильную классификацию всех обучающих примеров. gamma определяет, какое влияние оказывает каждый обучающий пример. Комбинация параметров была подобрана с помощью кросс-проверки, и была выбрана та, которая проявила на ней себя лучше остальных. Обучение проводилось на тренировочной подвыборке задачи, accuracy измерялась на калибровочной подвыборке. Лучшая комбинация параметров была выбрана с помощью поиска по сетке для C ∈ {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}, gamma ∈ {0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1} и kernel ∈ {rbf, linear, poly(deg=3)}. Графики 7 и 8 показывают зависимость точности классификации от параметров метода (графики с

остальными значениями коэффициентов регулязации представлены в приложении 1).

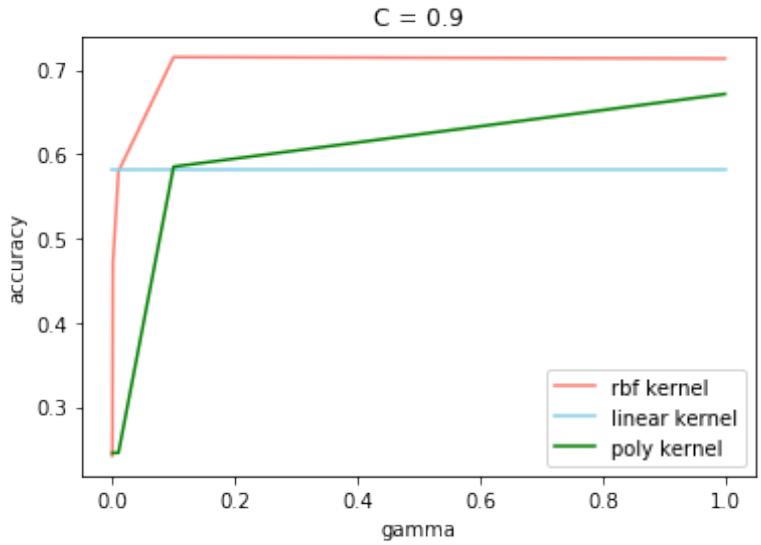


График 7. Зависимость точности от параметров метода опорных векторов.

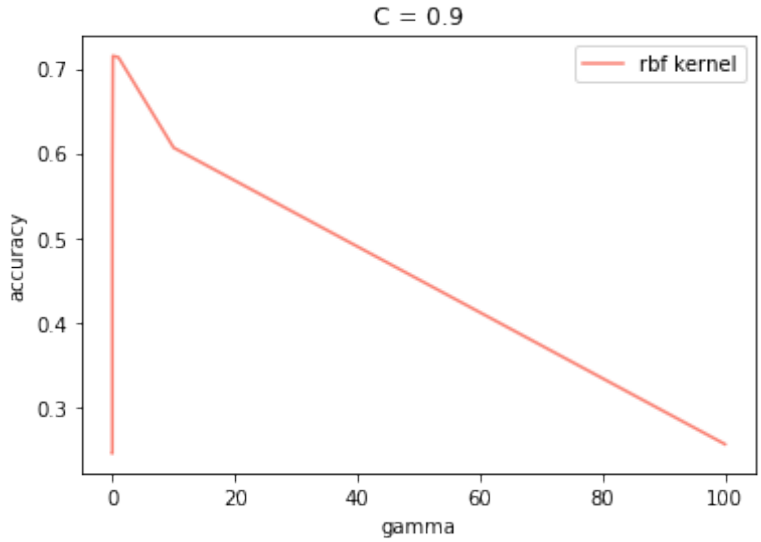


График 8. Зависимость точности от параметров метода опорных векторов (дальнейшее поведение с увеличение gamma для оптимальных параметров).

Самую высокую точность (71% на калибровочной выборке) модель достигла при использовании rbf ядра (с коэффициентом 0.1) и параметром регулязации равным 0.9. Финальная модель, используемая для установления фаций, обучается затем на тренировочной выборке задачи определения фаций с использованием выбранных параметров.

5.3 Наивный Байес

Суть метода заключается в определении наиболее вероятного класса. Алгоритм основан на теореме Байеса (3) с допущением о независимости признаков (то есть наличие признака в класса никак не связано с наличием других признаков). В работе был рассмотрен гауссовский случай классификатора (4).

$$P(y|x_1,\dots,x_n)=\frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)} \quad (3)$$

$$P(x_i|y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}e^{\frac{-(x_i-\mu_y)^2}{2\sigma_y^2}} \quad (4)$$

Где μ - математическое ожидание, σ^2 - дисперсия.

Более подробное метод описан в [14]. Финальная модель, используемая для установления фаций, обучается затем на тренировочной выборке задачи определения фаций с использованием выбранных параметров.

5.4. Деревья решений

Дерево решений позволяет определять принадлежность к тому или иному классу путем перехода по дереву, начиная от корневого узла, в направлении к листовым узлам. Каждый этап перехода осуществляется с учетом результатов проверки некоторых условий. Результат оценки всегда соответствует только одному из ребер, исходящих из узла принятия решений. Во время перехода прослеживается ребро, ведущее к следующему узлу принятия решений, в котором этот процесс повторяется. Переход по дереву прекращается после достижения его листа, соответствующего одному из классов. Более подробно метод описан в [8].

Эффективность дерева в первую очередь зависит от максимальных глубины (max_depth) и используемых на каждом разбиении признаков(max_features). Для подбора оптимальных параметров для каждой пары значений max_depth и max_features на тренировочной подвыборке

тренировочного набора задачи определения фаций была проведена кросс-валидация и была выбрана их лучшая комбинация. Зависимость точности классификации от параметров метода представлена на графике 9.

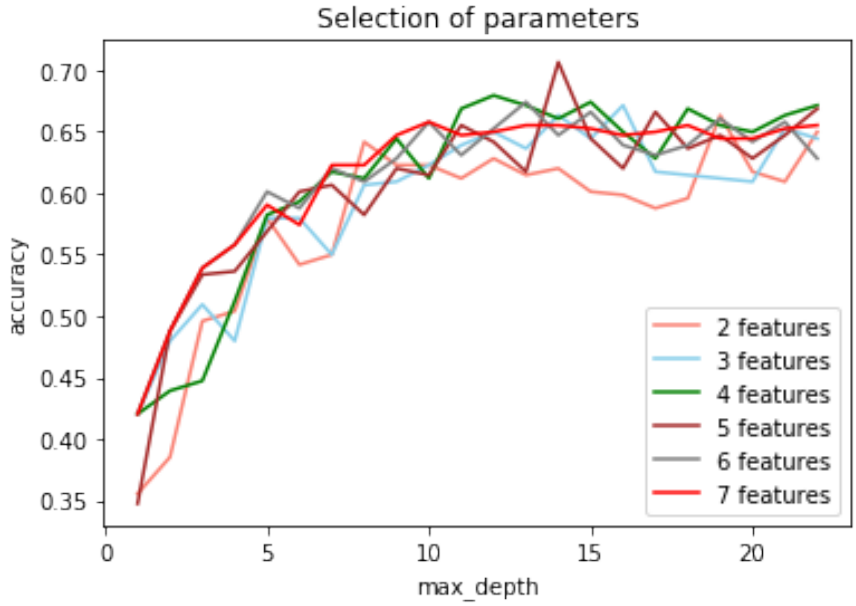


График 9. Зависимость точности от параметров дерева решений.

Самую высокую точность (72% на калибровочной выборке) модель достигла при параметрах max_depth = 14 и max_features = 5. Финальная модель, используемая для установления фаций, обучается затем на тренировочной выборке задачи определения фаций с использованием выбранных параметров.

5.5. Метод ближайших соседей

Идея метода ближайших соседей состоит в определении класса каждого нового объекта при помощи голосования его ближайших соседей (каждый сосед голосует за свой класс). То есть новому объекту присваивается класс данных, который имеет наибольшее количество представителей в ближайших соседях точки. Более подробно метод описан в [6].

Эффективность работы метода ближайших соседей зависит от выбора числа соседей и выбора метрики, измеряющей расстояние. Были рассмотрены евклидова, манхэттенская метрика и метрика Чебышева

(формулы 5-7). Комбинация параметров была подобрана с помощью кросс-проверки, и была выбрана та, которая проявила на ней себя лучше других.

$$dist = \sqrt{\sum (x-y)^2} \quad (5)$$

$$dist = \sum |x-y| \quad (6)$$

$$dist = max|x-y| \quad (7)$$

Обучения проводилось на тренировочной подвыборке задачи, accuracy измерялась на калибровочной подвыборке. Зависимость точности классификации от параметров представлена на графике 10.

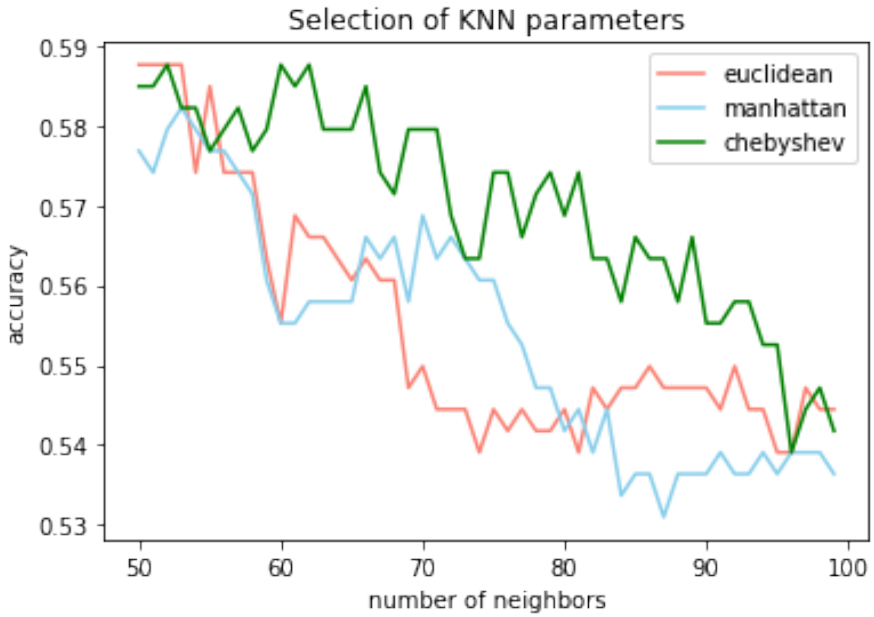


График 10. Зависимость точности от параметров метода ближайших соседей.

Самую высокую точность на калибровочной подвыборке метод показал при 53 соседях с расчетом расстояния с помощью евклидовой метрики. Финальная модель, используемая для установления фаций, обучается затем на тренировочной выборке задачи определения фаций с использованием выбранных параметров.

5.6. Результаты

Настроенные модели были обучены на тренировочной выборке задачи определения фаций по её геофизическим признакам и применены к тестовой скважине. Качество распознания фаций было определено с помощью метрики accuracy. В таблице 6 и на графике 11 представлено сравнение точности определения фаций для различных способов предобработки данных: заполнения неизмеренных значений нулями (zeros), с помощью алгоритма случайного леса (forest) и удалению значения фотоэффекта из выборок (nv). Это сравнение было проведено, ибо при неудачном выборе метода заполнения неизмеренных значений или при плохом тренировочном наборе такой способ заполнения пропусков может не только не улучшить, но и значительно ухудшить результаты. С помощью сравнения можно понять целесообразность применения методов машинного обучения для устранения отсутствующих значений в задаче каротажа.

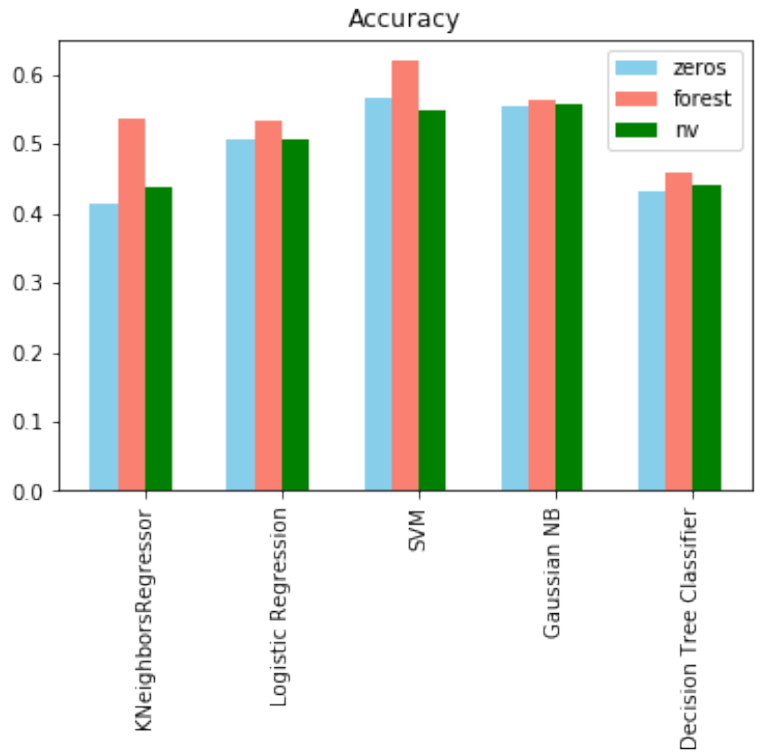


График 11. Accuracy для различных способов заполнения пропусков.

Для всех рассматриваемых методов машинного обучения набор данных со

значениями фотоэффекта, заполненными с помощью алгоритма случайного леса, получил лучший показатель accuracy.

Таблица 6. Accuracy для различных способах заполнения пропусков.

	нули	случайный лес	без PE
Метод ближайших соседей	0,41	0,54	0,44
Логистическая регрессия	0,51	0,53	0,5
Метод опорных векторов	0,56	**0,61**	0,55
Наивный Байес	0,55	0,56	0,56
Дерево решений	0,43	0,46	0,44

Самую высокую точность для задачи определений фаций по её геофизическим признакам (многоклассовая классификация, 9 видов фаций) показал метод опорных векторов (61%). Этот результат можно попытаться улучшить, учитывая особенности залегания фаций.

Accuracy с учетом «смежных» фаций в среднем составила 98% для всех методов, что говорит о практически полном отсутствии грубых ошибок при определении фаций.

Показатель accuracy для определения фаций в других скважинах варьируются от 50% до 61%. Причем результаты определения фаций в требующих восстановления значений фотоэффекта скважинах оказались не хуже, чем в не требующих подобного восстановления.

Без предварительно обработки данных и настройки методов МО accuracy составила в среднем составила 28% для набора данных с заменой пропусков нулями, 34% для наборов данных с заменой пропусков с помощью случайного леса.

В некоторых случаях при разработке скважин заказчиков может интересовать не точность определения всех фаций в скважине в целом, а точность определения какой-то одной из них (задача бинарной классификации). Для метода опорных векторов был вычислен показатель

accuracy (формула 8) для данной задачи (таблица 7).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (8)$$

Таблица 7. Accuracy для всех фаций.

	SS	CSiS	FSiS	SiSh	MS	WS	D	PS	BS
Метод опорных векторов	0.99	0.91	0.91	0.94	0.88	0.83	0.95	0.86	0.98

Также для нахождения наиболее эффективного алгоритма обнаружения каждой из фаций была использована F-мера. На графиках 12-20 представлены её значения по фациям.

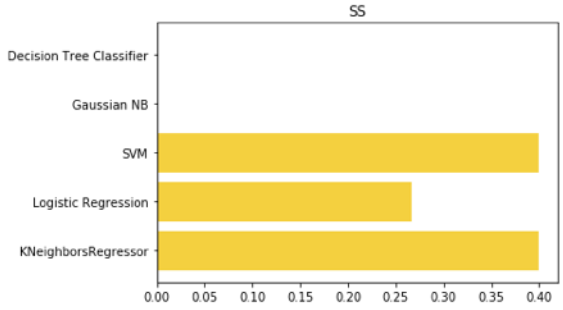


График 12. F-мера для неморского песчаника.

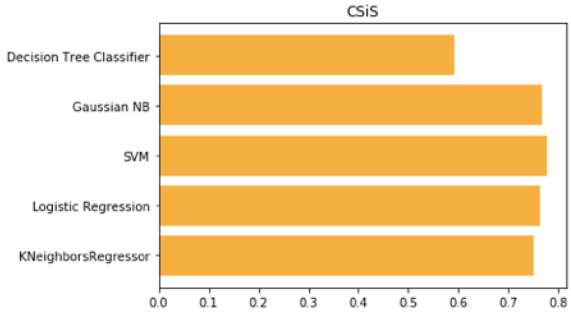


График 13. F-мера для неморского крупного алевролита.

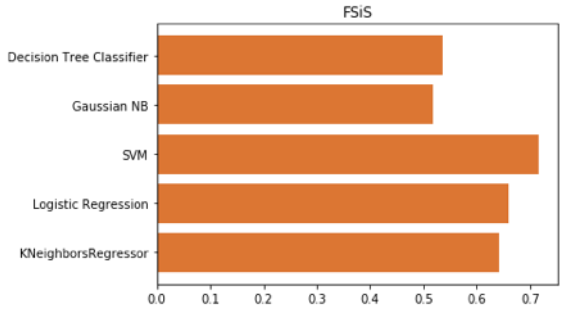


График 14. F-мера для неморского мелкого алевролита.

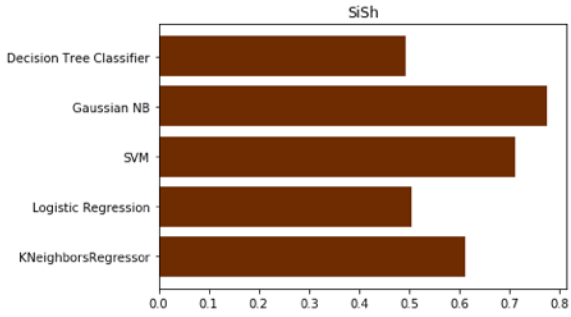


График 15. F-мера для морских алевролита и сланца.

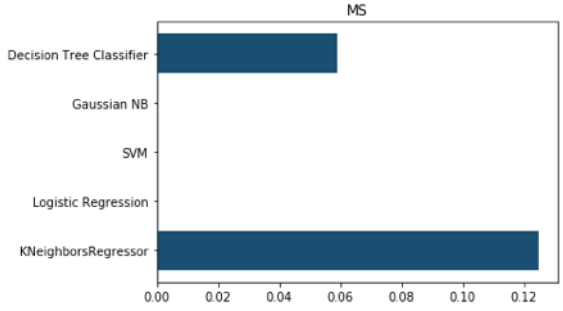


График 16. F-мера для аргиллита.

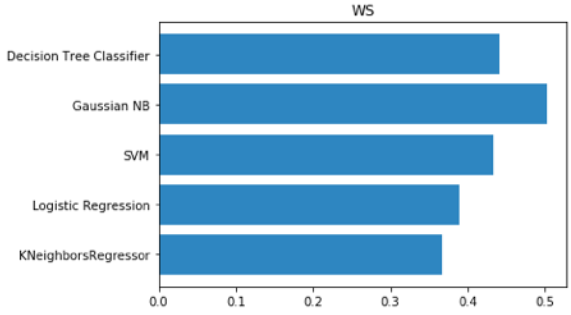


График 17. F-мера для ваккита.

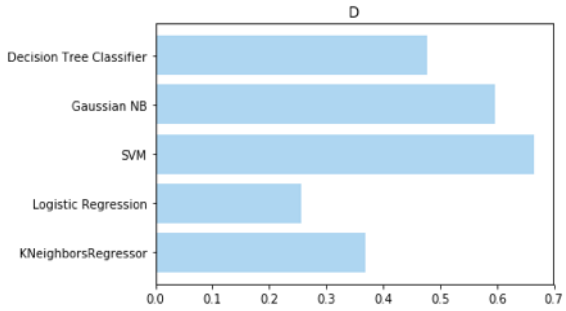


График 18. F-мера для неморского доломита.

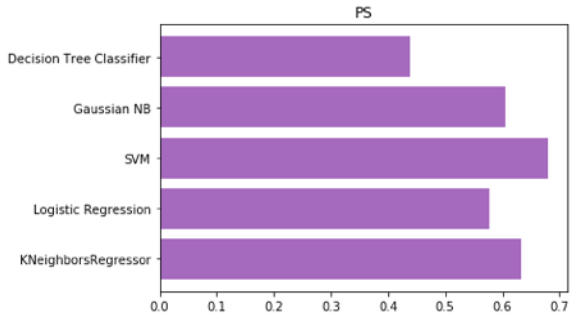


График 19. F-мера для пакстоуна-грейнстоуна.

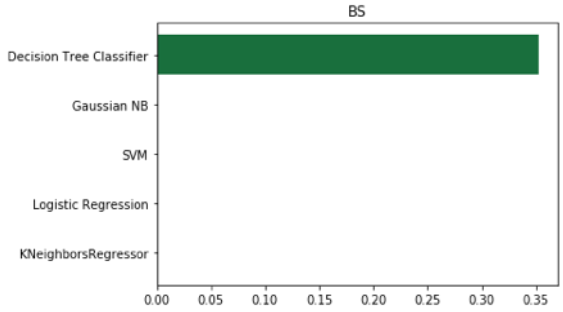


График 20. F-мера для phylloid-algal bafflestone.

Из графиков 12-20 видно, что метод опорных векторов хоть и показывает лучшие результаты при распознании фаций в скважине в целом, но не является оптимальным при определении каждой из фаций отдельно или вообще её не определяет. Так например, для обнаружения морских алевролита и сланца, а также ваккита целесообразнее использовать алгоритм наивный Байес, для определения аргиллита больше подошел метод ближайших соседей, а для phylloid-algal bafflestone - решающее дерево. Также стоит отметить, что алевролиты определились лучше всех фаций, их F-мера составила от 0,72 до 0,78. F-мера морского алевролита и сланца превзошла остальные и составила 0,78, также хорошо распознался доломит (0,67).

6. Постобработка данных

Для уточнения прогнозирования фаций было допущено предположение, что фации в соседних слоях коррелируют друг с другом.

Обработка уже полученных значений фаций на тестовой скважине заключалась в следующем: при обнаружении некоторого количества «изолированных» фаций, происходило их переопределение на наиболее широко представленные фации соседних слоев.

Эффективное количество «изолированных» фаций было определенно экспериментально (по обучающей выборке) и оказалось равно 2. Действительно, очень редко встречаются фации, занимающие лишь 1 фут.

На рисунке 3 представлена каротажная диаграмма с определенными по её значениям фациями с помощью метода опорных векторов, а на рисунке 4 показаны они же после предобработки. Точность распознавания фаций выросла на 2% и составила 63%.

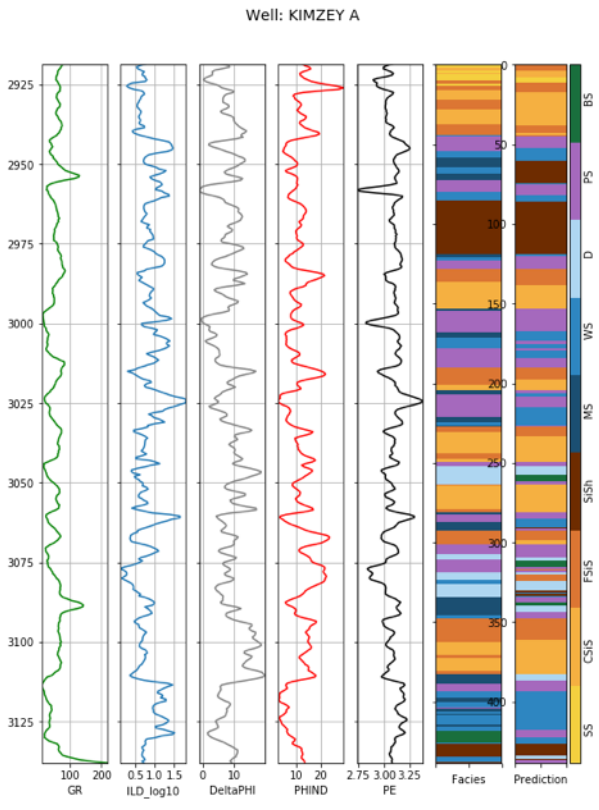


Рисунок 3. До предобработки

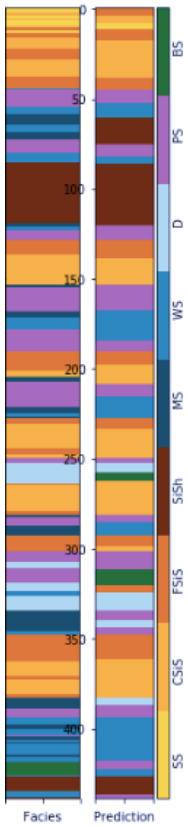


Рисунок 4. После предобработки

Постобработка была произведена для всех рассматриваемых методов машинного обучения (приложение 2) и во всех случая accuracy выросло на 1-3%.

Заключение

В работе была поставлена задача нахождения эффективного способа определения фаций, на основании данных геофизических исследований (данные каротажа). Для её решения был выполнен ряд подзадач.

Во-первых, был найден набор результатов каротажных исследований.

Во-вторых, так как для определения фаций планировалось использовать машинное обучение, данные были предварительно обработаны для корректной и эффективной работы алгоритмов. Также в ходе изучения исходных данных возникла вспомогательная подзадача восстановления неизмеренных значений фотоэффекта. Для её решения были использованы следующие методы машинного обучения: метод наименьших квадратов, случайный лес, логистическа регрессия, метод ближайших соседей, метод опорных векторов, и опорный вектор регрессии. Значения фотоэффекта были восстановлены с помощью алгоритма случайного леса, показавшего наилучший результат.

В-третьих, было произведено определение фаций. Для этого использовались несколько методов машинного обучения: логистическая регрессия, метод опорных векторов, наивный Байес, деревья решений и метод ближайших соседей, из которых лучшим оказался метод опорных векторов.

Также в работе был предложен способ постобработки определенных фаций, основанный на особенностях предметной области и повышающий точность классификации на 1-3%.

Самую высокую точность классификации фаций в скважине (9 видов фаций), равную 61%, показал метод опорных векторов. После постобработки этот результат вырос на 2% и составил 63%. Таким образом, поставленную в работе цель можно считать частично достигнутой. Результаты классификации фаций в скважине оказались весьма невысокими. Это произошло из-за довольно маленького числа исходных данных и отсутствия специальных геологических знаний.

В дальнейшем результаты можно улучшить. Положительно повлиять на результат могут более подробные геологические сведения об особенностях самих фаций, закономерностях их залегания.

Исходя из полученных в работе результатов и предположений о их возможном улучшении, можно утверждать об оправданности использования и дальнейшего изучения методов машинного обучения для определения фаций с помощью каротажных измерений.

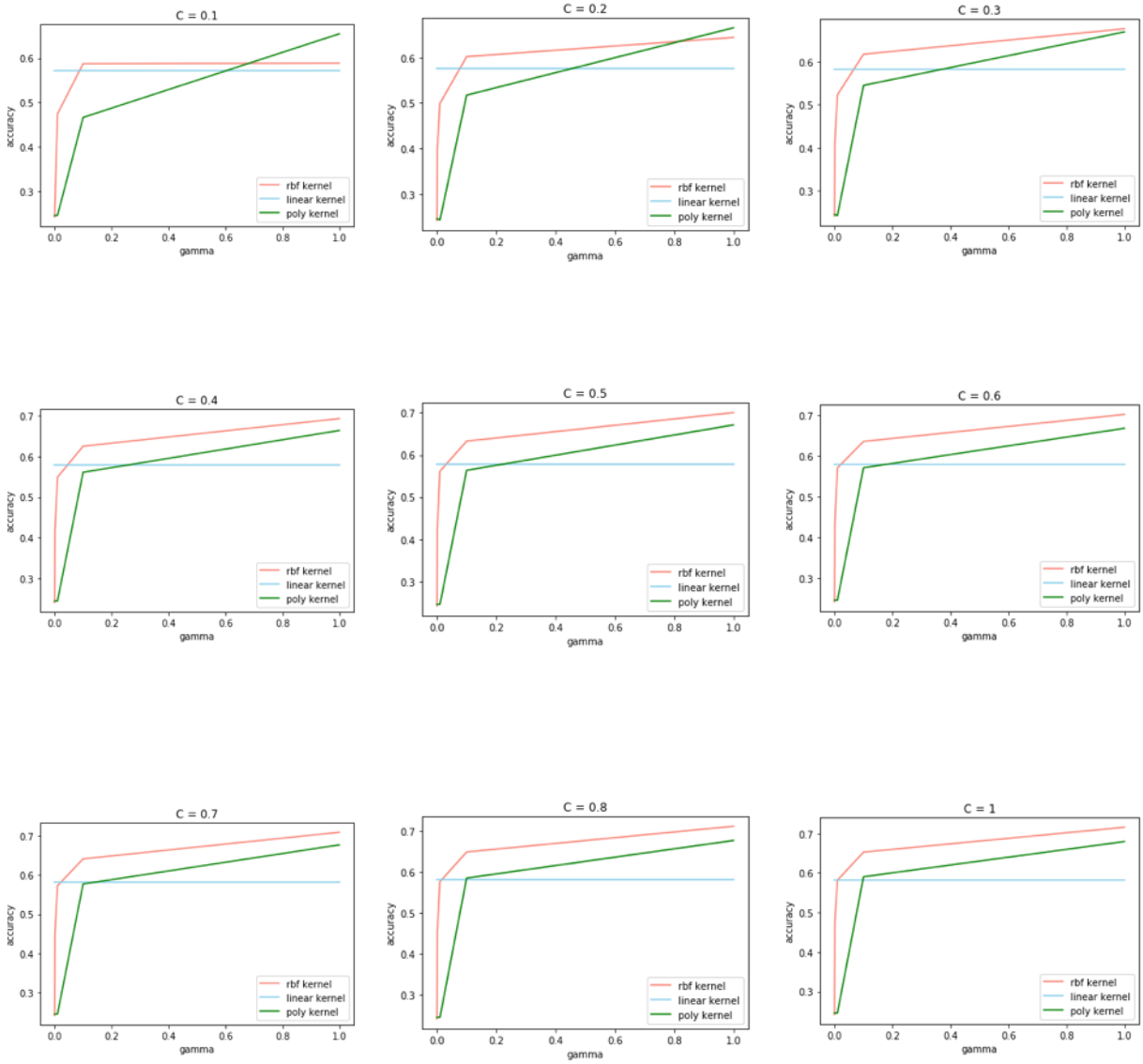
Список литературы

1. Давыдов В.А., Давыдов А.В. Очистка геофизических данных от шумов с использованием преобразования Гильберта-Хуанга // Актуальные инновационные исследования: наука и практика, №1, 2010.
2. Косков В. Н. Геофизические исследования скважин: учеб. пособие // Перм. гос. техн. ун-т. , Пермь, 2005. 317 с.
3. Мухамедиев Р.И. Методы машинного обучения в задачах геофизических исследований // Рига, 2016. 200 с.
4. Панов С.В, Парушкин М.Д., Фомин Ю.Н., Семибаламут В.М. Применение метода эмпирической модовой декомпозиции в обработке литосферных деформаций // Фундаментальные и прикладные вопросы горных наук, том 4, №3, 2017
5. Bestagini P., Lipari V., Tubaro S. A. Machine Learning Approach to Facies Classification Using Well Logs // SEG Technical Program Expanded Abstracts 2017, pp. 2137-2142.
6. Bishop C.M. Pattern Recognition and Machine Learning // New York, NY, USA: Springer; 2007. 738 pp.
7. Bremen L., Random Forests // Machine Learning, 45(1), 2001. pp. 5-32.
8. Decision Trees // URL: http://scikit-learn.org/stable/modules/tree.html#mathematical-formulation (online; accessed: 18.04.18).
9. Generalized Linear Models // URL: http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (online; accessed: 18.04.18).
10. Guyader A., Hengartner N. On the Mutual Nearest Neighbors Estimate in Regression // Journal of Machine Learning Research 14, 2013. pp. 2361-2376.
11. Huang N.E., Shen S.S.P. The Hilbert–Huang Transform and Its Applications // World Scientific Publishing Co. Pte. Ltd. 5 Toh Tuck. Link, Singapore
12. Li P., Zhang Y., Facies Characterization of a Reservoir in the North Sea Using Machine Learning Techniques // URL: http://cs229.stanford.edu/proj2016/report/LiZhang-FaciesCharacterizationOfAReservoirInTheNorthSeaUsingMachineLearningTechniques-Report.pdf (online; accessed: 11.03.18).

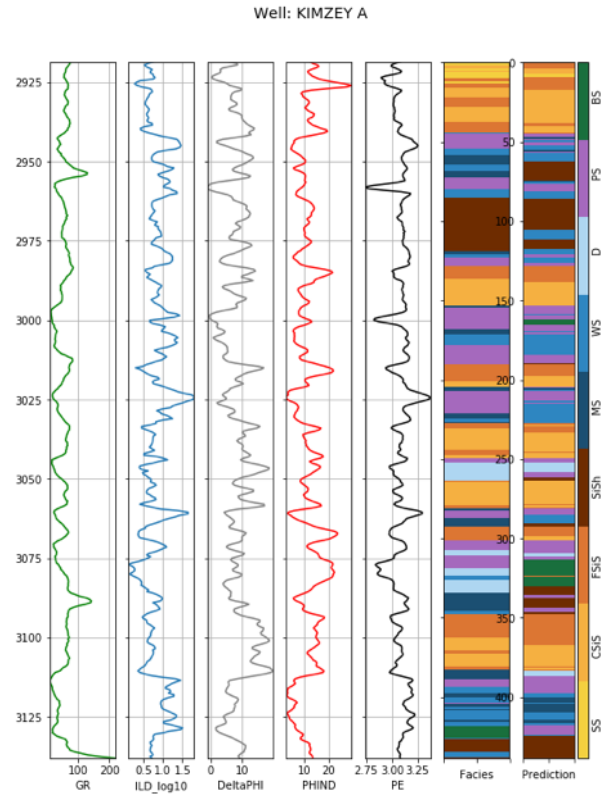
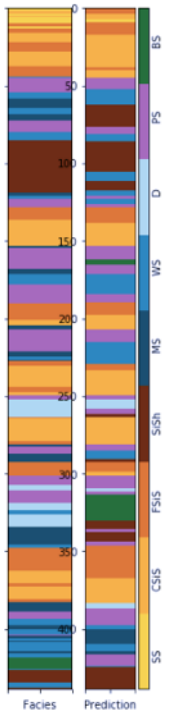
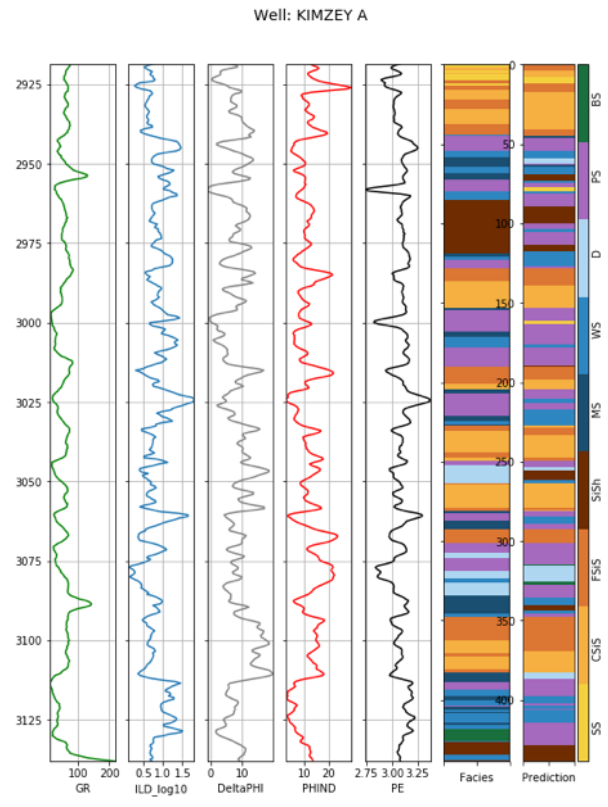
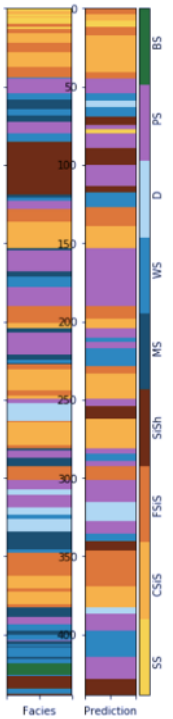
13. Moradi S., Motafakkerfard R., Riahi M.A., Sebtosheikh M.A. Separating Well Log Data to Train Support Vector Machines for Lithology. Prediction in a Heterogeneous Carbonate Reservoir // Iranian Journal of Oil & Gas Science and Technology, Vol. 4 (2015), No. 2, pp. 01-14.
14. Naive Bayes // URL: http://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes (online; accessed: 18.04.18).
15. Ng Andrew. Stanford CS229 Lecture Notes // URL: http://cs229.stanford.edu/notes/cs229-notes1.pdf (online; accessed: 10.04.18).
16. Schölkopf B., Smola A.J. A tutorial on support vector regression // Statistics and Computing 14, 2004. pp. 199–222.
17. Support Vector Machines // URL: http://scikit-learn.org/stable/modules/svm.html#svc (online; accessed: 18.04.18).

Приложение 1

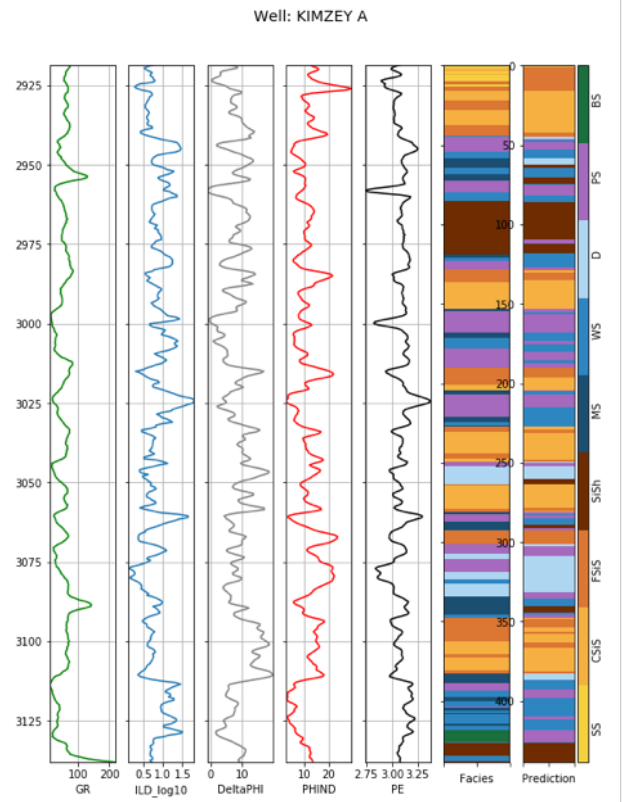
Зависимость accuracy от разных значений коэффициента gamma и параметра регуляризации для метода опорных векторов.

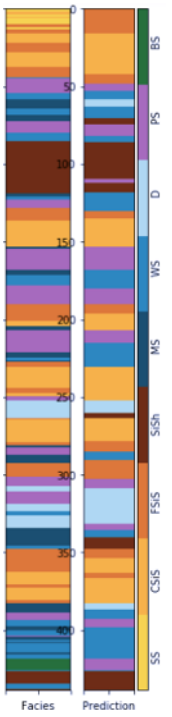


Приложение 2

до постобработки	после постобработки
Метод ближайших соседей	
	
Логистическая регрессия	
	

Наивный Байес





Дерево решений