УДК 622.27: 519.87
DOI 10.34822/1999-7604-2020-1-22-34

# МЕТОДЫ ОТБОРА ЗНАЧИМЫХ ПОКАЗАТЕЛЕЙ В МОДЕЛИ ОЦЕНКИ ЭФФЕКТИВНОСТИ ЭКСПЛУАТАЦИИ СКВАЖИН

**А. Ю. Вирстюк, А. А. Егоров** ✉
*Сургутский государственный университет, Сургут, Россия*
✉*E-mail: eaafit@gmail.com*

В статье представлены результаты отбора геолого-физических и технологических показателей эксплуатации нефтяных скважин в регрессионную модель оценки эффективности их работы. Рассмотрены методы прямого, обратного отбора переменных и метод на основе случайного леса. Выявлены достоинства и недостатки применения методов отбора в исследуемой предметной области, проведен сравнительный анализ на основе оценки точности моделей от количества учитываемых показателей. Подтверждена адекватность полученных регрессионных моделей.

*Ключевые слова:* нефтяная скважина, методы отбора, регрессионная модель.

# METHODS OF SELECTING SIGNIFICANT FEATURES IN A MODEL FOR ASSESSING THE EFFICIENCY OF WELL OPERATIONS

**A. Yu. Virstyuk, A. A. Egorov** ✉
*Surgut State University, Surgut, Russia*
✉*E-mail: eaafit@gmail.com*

The article presents the results of geological, material and technological features selection of oil well's efficiency for a regression model that evaluatestheir productivity. The methods of the forward, backward selection of variables and the method based on a random forest are considered. The advantages and disadvantages of the application of these methods in the subject field are revealed, and a comparative analysis is carried out based on an assessment of the model's accuracy according to the number of the selected features. The adequacy of the obtained regression models is confirmed.

*Keywords:* oil well, selection methods, regression model.

**Введение.** При оценке эксплуатации нефтяных скважин требуется анализ геолого-физических, технологических, химических и др. показателей. В связи с ростом трудноизвлекаемых залежей, повышением парафинистости, вязкости нефти и т. п. (иными словами, в связи с усложнением процесса нефтедобычи) число исходных данных (показателей) о работе нефтяных скважин неуклонно возрастает и требует перестройки или усложнения существующих гидродинамических, адаптационно-геологических и прочих моделей [1–2].

Ключевыми особенностями геологических процессов, составляющих основу нефтедобычи, являются невозможность их изучения в лабораторных условиях; зависимость природных систем от большого числа факторов; необходимость исследования значительного числа представителей изучаемого процесса для определения его свойств и закономерностей [3]. Все это требует оценки широкого спектра показателей. Но большое число показателей требует и большого числа наблюдений, чтобы охватить даже небольшую часть возможных конфигураций данных.

Увеличение объема выборки приводит к проблеме сложности нахождения линейных и нелинейных зависимостей, так как увеличивается влияние других факторов.

Для снижения размерности признакового пространства, в данном случае показателей эффективности работы нефтяных скважин, можно воспользоваться алгоритмами снижения размерности либо с помощью выделения признаков (feature extraction), либо с помощью отбора признаков (feature selection).

Алгоритмы выделения признаков (например, метод главных компонент – Principal Component Analysis) генерируют новые признаки, которые с трудом поддаются интерпретации. Исследуемая предметная область касается работы нефтяных скважин, что характеризуется сложностью и наличием большого числа нетривиальных закономерностей. Введение новых, плохо интерпретируемых признаков приводит к ещё большему её усложнению.

Методы отбора признаков, в свою очередь, помогают снизить размерность путем выделения информативных признаков и удаления менее полезных. Применение данных методов позволяет получить хорошо интерпретируемую регрессионную модель без введения новых признаков. Получаемая при этом модель будет характеризоваться высокой точностью, поскольку отобранные показатели хорошо коррелируются с результирующей переменной, что, в свою очередь, снижает вероятность переобучения модели (*overfitting*), при котором построенная модель хорошо объясняет примеры из тренировочной выборки, но относительно плохо работает с примерами тестовой выборки.

**Обзор методов отбора признаков.** В настоящее время можно выделить следующие классы методов отбора переменных:

1. Методы фильтрации (Filter-based).
2. Методы «обертки» (Wrapper-based).
3. Встроенные методы (Embedded).
4. Методы на основе случайных лесов (Random Forest).

Возможно применение ансамбля методов. Тогда говорят о гибридном методе отбора показателей.

Методы-фильтры не требуют привлечения алгоритмов обучения, они основаны на статистических методах. К примеру, для отбора признаков можно рассчитать дисперсию – средний квадрат отклонений индивидуальных значений признака ($x_i$) от средней величины ($\bar{x}$) [4], – рассчитываемую по формуле (1):

$$D = \frac{\sum_1^n (x_i - \bar{x})^2}{n}, \quad (1)$$

где $n$ – количество наблюдений.

Считается, что признаки с почти нулевой дисперсией не являются значимыми и их можно удалить.

Отбор показателей на основе методов-фильтров может также производиться на основе расчета критериев Фишера, хи-квадрата, построения матрицы корреляций и т. п.

Данные методы работают быстрее методов «обертки» и встроенных методов, поскольку их скорость линейно зависит от количества показателей, но они рассматривают каждую переменную изолированно. Из-за этого найти топ-$N$ наиболее коррелирующих переменных вообще не означает получить подмножество, на котором точность предсказания будет наивысшей [5].

Методы «обертки» фактически можно назвать методами перебора. Несмотря на то, что данные методы являются более надежными, чем остальные, они являются вычислительно самыми сложными.

Алгоритм перебора заключается в следующем: фиксируется небольшое число $N$, перебираются все комбинации по $N$-показателям. Выбирается лучшая из них, затем перебираются комбинации из $N+1$ показателей так, что предыдущая лучшая комбинация зафиксирована и перебирается только новый признак. Перебор происходит до тех пор, пока не будет достигнуто максимально допустимое число переменных или пока качество модели не перестанет значимо расти. Этот метод отбора переменных называется «прямой отбор» (Forward Selection).

Этот же алгоритм можно развернуть: зафиксировать полное пространство признаков и удалять признаки по одному. Так реализуется метод обратного исключения (Backward Selection).

Существует модификация метода прямого отбора – метод последовательного отбора (Stepwise Selection), заключающаяся в том, что на каждом шаге после включения новой переменной в модель осуществляется проверка на значимость остальных переменных, которые вошли в нее ранее.

Основным недостатком методов «обертки», как отмечалось ранее, является вычислительная сложность. Существует и еще один недостаток: в случае большого количества показателей и небольшого объема тренировочных данных возможно переобучение модели.

Встроенные методы производят отбор признаков внутри процесса расчета модели. По вычислительной сложности данные методы занимают промежуточное положение: они медленнее методов фильтрации, но быстрее методов «обертки».

Встроенные методы опираются на понятие «регуляризация», идея которой заключается в том, чтобы построить алгоритм, минимизирующий не только среднеквадратичную ошибку, но и количество используемых переменных. К встроенным методам относятся: LASSO, Ridge Regression и т. п.

В отдельный класс методов отбора переменных можно отнести методы на основе случайного леса – Random Forest.

Методы на основе случайного леса осуществляют построение большого числа (ансамбля) деревьев решений, каждое из которых строится по выборке, получаемой из исходной обучающей выборки с помощью BootStrap (выборка с возвращением) [6].

Для отбора значимых показателей во время построения модели для каждого элемента обучающей выборки производится расчет ошибки out-of-bag (далее – ООВ). Поскольку применяется выборка с возвращением, то некоторые объекты могут выбираться несколько раз, а некоторые – вообще ни разу. Для каждого объекта выставляется вес. Если копия одного и того же объекта попадает в BootStrap-выборку несколько раз, то и штраф за ошибку на этом объекте будет больше.

Описание алгоритма расчета ошибки ООВ представлено в работе [7], автором которой является основоположник концепции Random Forest Лео Брейман.

Затем для каждого объекта такая ошибка усредняется по всему случайному лесу. Важность признака оценивается путем усреднения по всем деревьям разности показателей ООВ до и после перемешивания значений.

Применение методов на основе случайного леса обеспечивает защиту от переобучения. Для их построения требуется задание минимального количества настраиваемых параметров. Работа методов характеризуются незначительными вычислительными затратами.

**Описание исходных данных.** Для проведения сравнительного анализа различных методов отбора переменных будет использоваться выборка, характеризующая работу 2 000 нефтяных скважин. Работа нефтяных скважин описывается 30 геолого-физическими и технологическими показателями. В табл. 1 представлены названия показателей и их единицы измерения.

*Таблица 1*

**Исходные данные для оценки эффективности работы нефтяных скважин**

| № | Показатель | Единицы измерения |
|---|---|---|
| 1 | Давление насыщения нефти газом | МПа |
| 2 | Пластовая температура | град. С |
| 3 | Среднее значение открытой пористости (коэффициента пористости) | доли ед. |
| 4 | Сжимаемость породы | $\frac{1}{\text{МПа}} * 10^{-4}$ |
| 5 | Плотность нефти в пластовых условиях | кг/м$^3$ |
| 6 | Вязкость нефти в пластовых условиях | МПа*с |

*Окончание табл. 1*

| № | Показатель | Единицы измерения |
|---|---|---|
| 7 | Сжимаемость нефти | $\frac{1}{\text{МПа}} * 10^{-4}$ |
| 8 | Содержание парафина в нефти | % |
| 9 | Содержание серы в нефти | % |
| 10 | Проницаемость | $*10^{-3}$мкм$^2$ |
| 11 | Коэффициент нефтенасыщенности | доли ед. |
| 12 | Нефтяная площадь залежи | тыс. м$^2$ |
| 13 | Эффективная толщина пласта | м |
| 14 | Общая толщина пласта | м |
| 15 | Содержание смол и асфальтенов | % |
| 16 | Коэффициент песчанистости | доли ед. |
| 17 | Количество влияющих нагнетательных скважин | шт. |
| 18 | Время работы скважины | сутки |
| 19 | Среднее расстояние между добывающей и нагнетательной скважиной | м |
| 20 | Средний радиус влияния | м |
| 21 | Фактическая приемистость влияющих нагнетательных скважин | м$^3$/сут |
| 22 | Среднее буферное давление влияющих нагнетательных скважин | мПа |
| 23 | Суммарный объем закачанной воды по интерферирующим нагнетательным скважинам | м$^3$ |
| 24 | Объем добываемой нефти | тонн |
| 25 | Объем добываемой воды | тонн |
| 26 | Объем добываемой жидкости | тонн |
| 27 | Коэффициент охвата пласта заводнением | % |
| 28 | Забойное давление | ат |
| 29 | Обводненность продукции | % |
| 30 | Коэффициент продуктивности | $^{\text{т}}/_{(\text{сут} \cdot \text{атм})}$ |

*Примечание:* составлено авторами.

Показатели имеют различную размерность, поэтому перед применением методов отбора показателей, была произведена их нормализация методом *Z*-масштабирования.

**Сравнительный анализ некоторых методов отбора показателей, характеризующих работу нефтяных скважин.** Ниже будет проведен сравнительный анализ основных методов «обертки» (прямой, обратный отбор переменных) и метода на основании случайного леса.

Для построения обучающихся моделей необходимо разбить исходное множество на две выборки: тренировочные данные, необходимые для построения самой модели, и тестовые данные для проверки качества построенной модели.

Для выполнения этой задачи можно воспользоваться встроенной *k*-блочной перекрестной проверкой (*k*-fold cross validation). В *k*-блочной перекрестной проверке данные разделяются на *k*-частей (блоков). Модель обучается на *k-1* блоках, представляющих один тренировочный набор. Обучение повторяется *k*-раз, на каждом шаге в качестве тестового набора используется другой блок. Результативность модели представляет собой среднее результативности на каждой из *k*-итераций. Соответственно, один блок используется для тестирования полученной модели [8]. Число итераций *k* в рамках данной работы равно 10.

Для оценки качества отбора показателей различными методами используется точность модели. Точность модели – это доля прогнозируемых положительных результатов, которые являются действительно верно-положительными результатами для всех положительно предсказанных объектов. Другими словами, точность дает нам ответ на вопрос: из всех объектов, которые классифицированы как принадлежащие классу, сколько на самом деле принадлежит ему? [9, 10].

**Метод прямого отбора показателей.** В качестве зависимого показателя был выбран коэффициент продуктивности ($^{\text{т}}/_{(\text{сут} \cdot \text{атм})}$), представленный в табл. 1 под номером 30. Остальные показатели считаются независимыми.

На рис. 1 показана зависимость точности построенной модели прямого отбора показателей от количества учитываемых переменных. Из графика видно, что наибольшее значение точность принимает при семи показателях, что показано выносной линией. Далее точность модели снижается.

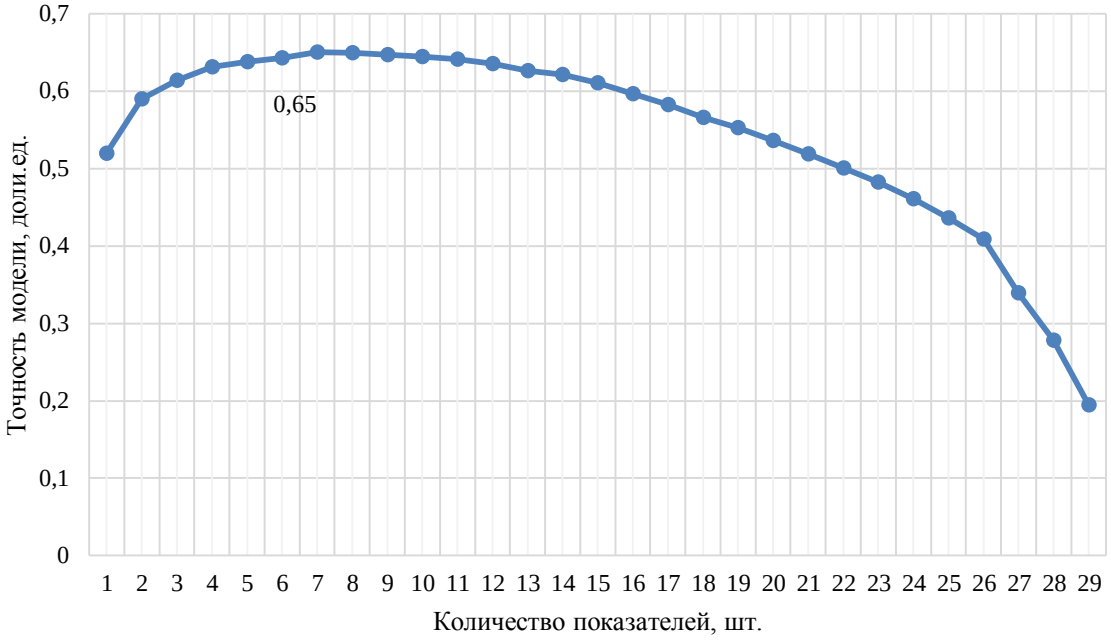


**Рис. 1. Зависимость точности модели от числа показателей для метода прямого отбора**
*Примечание:* составлено авторами.

В табл. 2 представлены названия отбираемых показателей и соответствующая им точность модели для первых 10 итераций.

Для удобства названия показателей из табл. 1 сокращены до двух слов.

*Таблица 2*

**Первые 10 итераций метода прямого отбора показателей**

| № | Показатели (из таблицы 1) | Точность модели |
|---|---|---|
| 1 | Давление насыщения | 0,519 |
| 2 | Давление насыщения, коэффициент пористости | 0,589 |
| 3 | Давление насыщения, коэффициент пористости, среднее расстояние | 0,613 |
| 4 | Давление насыщения, коэффициент пористости, среднее расстояние, коэффициент охвата | 0,631 |
| 5 | Давление насыщения, коэффициент пористости, содержание парафина, среднее расстояние, коэффициент охвата | 0,637 |
| 6 | Давление насыщения, коэффициент пористости, содержание парафина, среднее расстояние, коэффициент охвата | 0,643 |
| 7 | Давление насыщения, коэффициент пористости, плотность нефти, содержание парафина, время работы, среднее расстояние, коэффициент охвата | 0,650 |
| 8 | Давление насыщения, коэффициент пористости, плотность нефти, содержание парафина, время работы, среднее расстояние, объем воды, коэффициент охвата | 0,649 |
| 9 | Давление насыщения, коэффициент пористости, плотность нефти, содержание парафина, коэффициент песчанистости, время работы, среднее расстояние, объем воды, коэффициент охвата | 0,647 |
| 10 | Давление насыщения, коэффициент пористости, плотность нефти, содержание парафина, смолы и асфальтены, коэффициент песчанистости, время работы, среднее расстояние, объем воды, коэффициент охвата | 0,645 |

*Примечание:* составлено авторами.

Метод прямого отбора выделяет 7 показателей в качестве основных, четыре из которых описывают геолого-физические особенности скважин (давление насыщения, коэффициент пористости, плотность нефти, содержание парафина), а остальные – технологические (время работы, среднее расстояние между нефтяной и нагнетательными скважинами и коэффициент охвата пласта заводнением).

**Метод рекурсивного исключения признаков.** Алгоритм рекурсивного исключения выполняется следующим образом: вначале строится модель по всем предикторам, которые ранжируются по их важности; далее рассматривается последовательность подмножеств $S$ ($S_1 > S_2$, и т. д.) из переменных наивысшего ранга. На каждой итерации подмножества $S_i$ ранги предикторов пересматриваются, а модели пересчитываются. Итоговая модель основывается на подмножестве $S_i$, обеспечивающем оптимум заданному критерию качества [11].

На рис. 2 представлена зависимость точности построенной модели рекурсивного исключения признаков от количества учитываемых показателей. Видно, что максимальная точность модели достигается при 13 показателях.

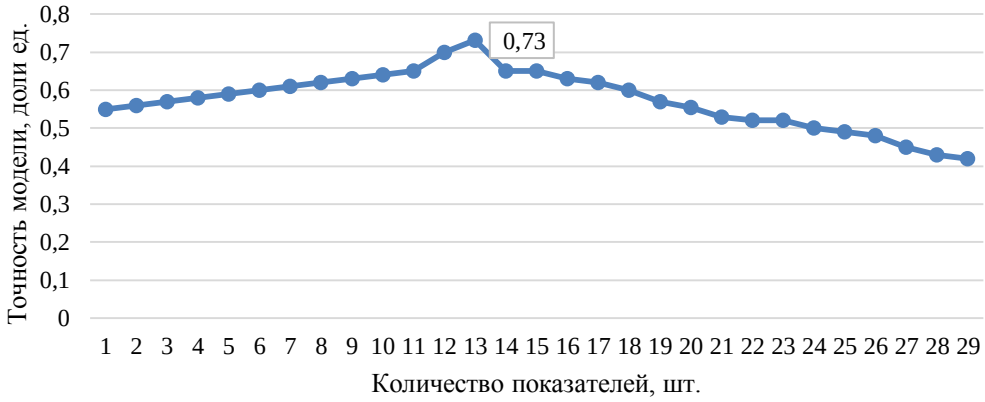


**Рис. 2. Зависимость точности модели от числа показателей для метода рекурсивного исключения признаков**
*Примечание:* составлено авторами.

Как отмечалось выше, данный метод каждому показателю ставит в соответствие ранг. Ранг, равный 1, показывает, что показатель необходимо оставить (табл. 3).

*Таблица 3*

**Ранги показателей, полученные методом рекурсивного исключения признаков**

| Показатели (из таблицы 1) | Ранг | Показатели (из таблицы 1) | Ранг |
|---|---|---|---|
| Давление насыщения нефти газом | 1 | Коэффициент песчанистости | 11 |
| Пластовая температура | 1 | Количество влияющих нагнетательных скважин | 10 |
| Среднее значение открытой пористости (коэффициента пористости) | 1 | Время работы скважины | 7 |
| Сжимаемость породы | 1 | Среднее расстояние между добывающей и нагнетательной скважинами | 13 |
| Плотность нефти в пластовых условиях | 1 | Средний радиус влияния | 5 |
| Вязкость нефти в пластовых условиях | 1 | Фактическая приемистость влияющих нагнетательных скважин | 14 |
| Сжимаемость нефти | 1 | Среднее буферное давление влияющих нагнетательных скважин | 6 |
| Содержание парафина в нефти | 1 | Суммарный объем закачанной воды по интерферирующим нагнетательным скважинам | 8 |
| Содержание серы в нефти | 1 | Объем добываемой нефти | 12 |

*Окончание табл. 3*

| Показатели (из таблицы 1) | Ранг | Показатели (из таблицы 1) | Ранг |
|---|---|---|---|
| Проницаемость | 1 | Объем добываемой воды | 4 |
| Коэффициент нефтенасыщенности | 15 | Объем добываемой жидкости | 9 |
| Нефтяная площадь залежи | 2 | Коэффициент охвата пласта заводнением | 1 |
| Эффективная толщина пласта | 1 | Забойное давление | 16 |
| Общая толщина пласта | 1 | Обводненность продукции | 17 |
| Содержание смол и асфальтенов | 3 | | |

*Примечание:* составлено авторами.

Таким образом, метод рекурсивного исключения признаков в качестве основных показателей предлагает оставить 13, причем 12 из них относятся к геолого-физической группе и только один – коэффициент охвата пласта заводнением – относится к технологической. Несмотря на достаточно высокое значение точности полученной модели, неравномерный отбор признаков, с уклоном на геолого-физическую составляющую, ставит под сомнение возможность применения данного метода к исследуемым данным.

**Метод на основе случайного леса.** Отбор показателей на основе метода случайного леса представляет собой альтернативное решение классическим статистическим методам.

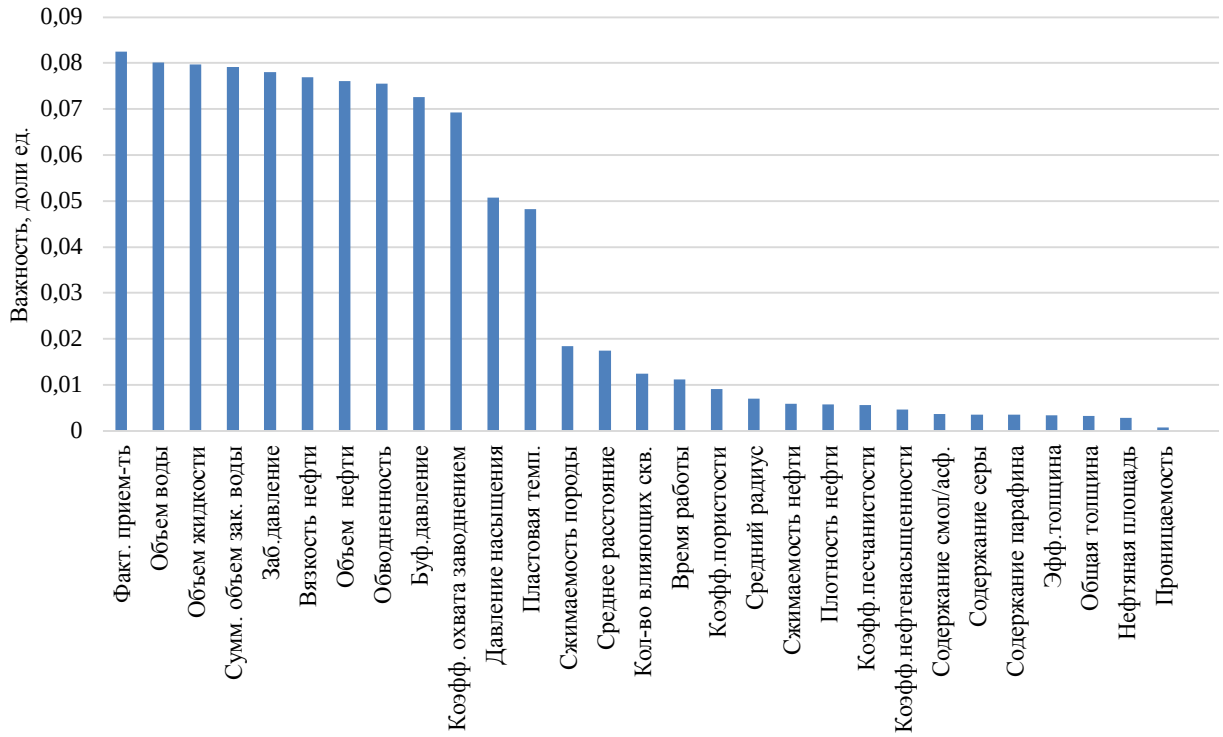


**Рис. 3. Важность показателей, полученная методом случайного леса**
*Примечание:* составлено авторами.

Поскольку распространенной проблемой деревьев принятия решений является то, что они очень плотно прилегают к тренировочным данным, это привело к появлению случайных лесов, представляющих собой множество деревьев принятия решений [8].

Входным параметром данного метода является задание количества деревьев решений для включения в лес. В рамках данной работы было задано 100 деревьев.

Метод основывается на вычислении относительной важности показателей. Чем выше число, тем важнее признак. В сумме все оценки важности составляют 1.

Из рис. 3 видно, что первые 10 показателей (фактическая приемистость, объемы нефти, воды и жидкости, суммарный объем закачанной воды, забойное давление, вязкость

нефти, обводненность, буферное давление, коэффициент охвата пласта заводнением) дают приблизительно по 8 % вклада каждый в результирующую переменную. Показатели «давление насыщения» и «пластовая температура» составляют приблизительно 5 % вклада каждый. Важность остальных показателей колеблется в пределах от 0 % до 2 %.

Для определения числа важных показателей был построен график зависимости точности модели от числа показателей, аналогичный графикам для ранее рассмотренных методов.

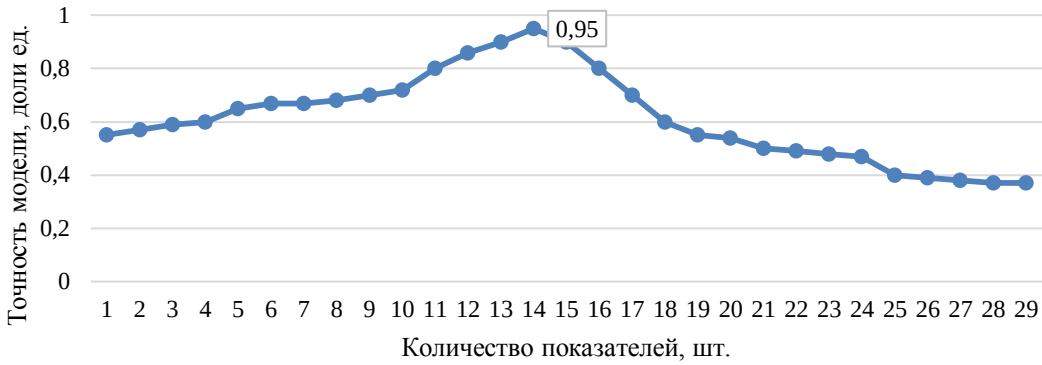


**Рис. 4. Зависимость точности модели от числа показателей для метода на основе случайных лесов**
*Примечание:* составлено авторами.

Из рис. 4 видно, что наибольшая точность достигается при введении в модель 14 показателей работы нефтяных скважин. Это подтверждают и рассчитанные важности переменных, представленные на рис. 3. Видно, что именно 14 показателей характеризуются высокими значениями важности, остальные дают небольшой вклад в результирующую переменную.

Из отобранных показателей к геолого-физической группе относятся: вязкость нефти, давление насыщения, пластовая температура, сжимаемость породы. К технологической группе относятся: фактическая приемистость влияющих нагнетательных скважин, объемы воды, жидкости и нефти, забойное и буферное давления, суммарный объем закачанной воды, обводненность, коэффициент охвата пласта заводнением, среднее расстояние между нефтяной и нагнетательными скважинами.

**Сравнение методов отбора признаков.** Ниже представлена результирующая таблица отобранных показателей каждым из трех рассмотренных методов (табл. 4).

*Таблица 4*

**Результирующая таблица отобранных показателей различными методами**

| Показатель (из таблицы 1) | Метод прямого отбора | Метод рекурсивного исключения | Метод случайных лесов |
|---|---|---|---|
| Давление насыщения нефти газом | + | + | + |
| Пластовая температура | - | + | + |
| Среднее значение открытой пористости (коэффициента пористости) | + | + | - |
| Сжимаемость породы | - | + | + |
| Плотность нефти в пластовых условиях | + | + | - |
| Вязкость нефти в пластовых условиях | - | + | + |
| Сжимаемость нефти | - | + | - |
| Содержание парафина в нефти | + | + | - |
| Содержание серы в нефти | - | + | - |
| Проницаемость | - | + | - |
| Коэффициент нефтенасыщенности | - | - | - |
| Нефтяная площадь залежи | - | | - |
| Эффективная толщина пласта | - | + | - |

*Окончание табл. 4*

| Показатель (из таблицы 1) | Метод прямого отбора | Метод рекурсивного исключения | Метод случайных лесов |
|---|---|---|---|
| Общая толщина пласта | - | + | - |
| Содержание смол и асфальтенов | - | - | - |
| Коэффициент песчанистости | - | - | - |
| Количество влияющих нагнетательных скважин | - | - | - |
| Время работы скважины | + | - | - |
| Среднее расстояние между добывающей и нагнетательной скважинами | + | - | + |
| Средний радиус влияния | - | - | - |
| Фактическая приемистость влияющих нагнетательных скважин | - | - | + |
| Среднее буферное давление влияющих нагнетательных скважин | - | - | + |
| Суммарный объем закачанной воды по интерферирующим нагнетательным скважинам | - | - | + |
| Объем добываемой нефти | - | - | + |
| Объем добываемой воды | - | - | + |
| Объем добываемой жидкости | - | - | + |
| Коэффициент охвата пласта заводнением | + | + | + |
| Забойное давление | - | - | + |
| Обводненность продукции | - | - | + |
| Число отобранных показателей | 7 | 13 | 14 |

*Примечание:* составлено авторами.

Из табл. 4 видно, что показатели, отобранные разными методами в качестве основных, тоже различны. Метод прямого отбора отличается небольшим числом отобранных показателей по сравнению с методами обратного отбора и на основе случайного леса. Метод обратного отбора характеризуется преобладанием геолого-физических показателей. Метод на основании случайного леса выделяет наибольшее число показателей, относящихся как к геолого-физической, так и к технологической группам.

На рис. 5 представлен общий график зависимости точности каждого из рассмотренных методов от числа показателей.

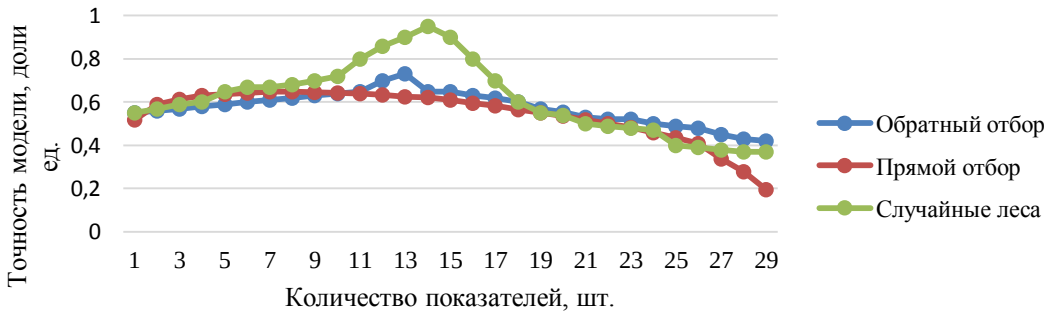


**Рис. 5. Зависимости точности модели от числа показателей для сравниваемых методов**
*Примечание:* составлено авторами.

Для того, чтобы определить, насколько качественно каждый из рассмотренных ранее методов отобрал показатели, были построены соответствующие уравнения множественной линейной регрессии. Выбор регрессионной модели подробно описывается в работах [12–13, 17]. В качестве зависимой переменной выступал «коэффициент продуктивности», а в качестве регрессоров – отобранные показатели.

Для каждого уравнения регрессии были рассчитаны коэффициенты множественной корреляции, коэффициент множественной детерминации и скорректированный коэффициент множественной детерминации.

Коэффициент множественной корреляции $R$ описывает тесноту линейной связи между зависимой переменной и группой других независимых показателей.

Коэффициент множественной детерминации $R^2$ показывает, какая доля вариации изучаемого результативного признака (в данном случае коэффициента продуктивности) объясняется влиянием факторов, включенных в уравнение множественной регрессии [15].

В свою очередь, скорректированный коэффициент множественной детерминации $R^2_{adj}$ позволяет сравнивать модели с различным числом регрессоров (показателей) [16, 17]. Расчет скорректированного коэффициента множественной детерминации осуществляется по формуле:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2), \quad (2)$$

где $n$ – число наблюдений;

$p$ – число факторов в уравнении регрессии.

В табл. 5 представлены значения коэффициента множественной корреляции, коэффициента множественной детерминации и скорректированного коэффициента множественной детерминации для трех уравнений регрессии с регрессорами, отобранными сравниваемыми методами.

*Таблица 5*

**Коэффициент множественный корреляции, коэффициент множественной детерминации и скорректированный коэффициент множественной детерминации уравнений линейной множественной регрессии**

| | **Метод прямого отбора** | **Метод рекурсивного исключения** | **Метод случайных лесов** |
|---|---|---|---|
| $R$ | 0,61 | 0,73 | 0,89 |
| $R^2$ | 0,37 | 0,53 | 0,79 |
| $R^2_{adj}$ | 0,36 | 0,52 | 0,78 |

*Примечание:* составлено авторами.

Таблица 5 показывает, что показатели, отобранные методом случайных лесов, лучше объясняют результирующую переменную – коэффициент продуктивности.

Для того, чтобы оценить статистическую значимость уравнения регрессии и полученного коэффициента множественной детерминации, необходимо воспользоваться $F$-критерием Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}, \quad (3)$$

где $p$ – число независимых переменных в уравнении регрессии;

$n$ – число наблюдений.

Табличное значение F-критерия Фишера при степенях свободы $f_1 = p = 7$ и $f_2 = n - p - 1 = 1992$ приблизительно равен 1,94. Для $f_1 = p = 13$ и $f_2 = n - p - 1 = 1986$ и $f_1 = p = 14$ и $f_2 = n - p - 1 = 1985$ и приблизительно равен 1,75.

Для каждого из уравнений линейной множественной регрессии табличное значение $F$-критерия Фишера намного меньше рассчитанного значения, соответственно, гипотезу о статистической незначимости уравнений регрессии можно отвергнуть.

На основании вышеизложенного можно сделать вывод, что уравнение регрессии, включающее показатели, отобранные методом на основе случайного леса, объясняет почти 80 % наблюдений.

Адекватность этой регрессионной модели подтверждает и нормальное распределение остатков. На рис. 6 видно, что остатки расположены близко к прямой, что позволяет сделать предположение о том, что остатки распределены по нормальному закону.

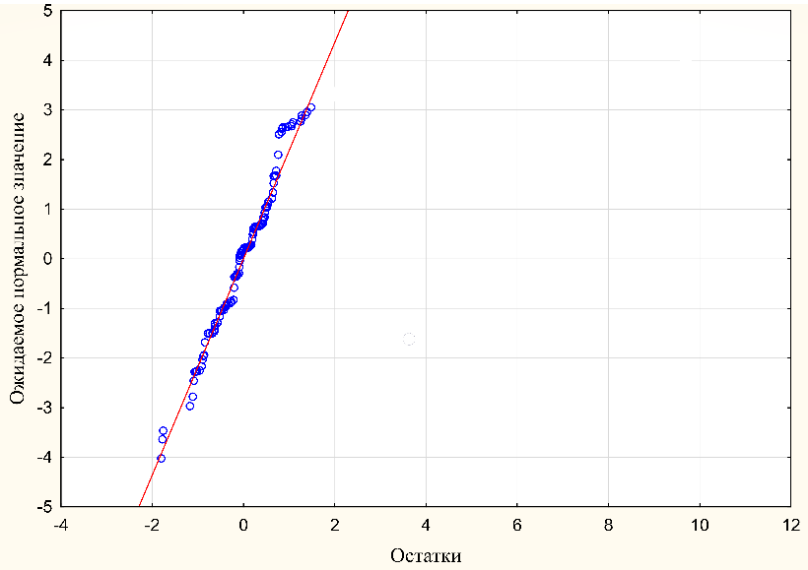


**Рис. 6. График остатков на нормальной вероятностной бумаге**
*Примечание:* составлено авторами.

Уравнение регрессии с показателями методов прямого и обратного отборов объясняет только 37 % и 53 % соответственно, что является недостаточным при прогнозировании эффективности работы нефтяных скважин.

Уравнение линейной множественной регрессии, описывающее зависимость коэффициента продуктивности ($y$) от геолого-физических и технологических показателей, отобранных методом на основе случайного леса, можно представить следующим образом:

$$\begin{aligned} Y = &- 3{,}9x_1 - 0{,}12x_2 + 1{,}2x_3 - 0{,}94x_4 + 0{,}45x_5 + 1{,}2x_6 + 1{,}79x_7 + 2{,}5x_8 + \\ &+ 0{,}71x_9 - 0{,}23x_{10} + 2{,}1x_{11} - 1{,}7x_{12} - 1{,}1x_{13} - 1{,}1x_{14}, \end{aligned} \quad (4)$$

где $x_1$ – давление насыщения;
$x_2$ – пластовая температура;
$x_3$ – сжимаемость породы;
$x_4$ – вязкость нефти;
$x_5$ – фактическая приемистость;
$x_6$ – среднее расстояние между скважинами;
$x_7$ – буферное давление;
$x_8$ – объем закачанной воды;
$x_9$ – объем добываемой нефти;
$x_{10}$ – объем добываемой воды;
$x_{11}$ – объем добываемой жидкости;
$x_{12}$ – коэффициент охвата пласта заводнением;
$x_{13}$ – забойное давление;
$x_{14}$ – обводненность продукции.

Возможно получение еще большего коэффициента множественной детерминации при построении уравнения линейной множественной регрессии на основе метода случайного леса. Для этого необходимо изменить входные показатели при построении деревьев случайного леса, например, изменить количество деревьев или их глубину.

**Заключение.** В данной статье был проведен сравнительный анализ трех методов отбора показателей, два из которых (метод прямого отбора и рекурсивного исключения признаков) относятся к методам «обертки», а третий является методом на основе случайного леса. Точность полученной модели самая высокая у метода на основе случайного леса: она составляет 95 %.

При построении уравнения линейной множественной регрессии на основании отобранных показателей каждым методом наибольший коэффициент детерминации был также получен для метода на основе случайного леса – 0,79. Для методов прямого отбора и рекурсивного исключения признаков коэффициент детерминации составил 0,37 и 0,53 соответственно.

Статистическая значимость уравнения регрессии и коэффициента детерминации была проверена с помощью *F*-критерия Фишера.

Полученные результаты подтверждают более высокую эффективность метода на основе случайного леса по отбору геолого-физических и технологических показателей работы нефтяных скважин по сравнению с классическими методами «обертки».

Следует отметить, что одним из ключевых недостатков метода на основе случайного леса является большой размер получающихся моделей, напрямую зависящий от количества деревьев случайного леса, но высокая точность предсказания делает данный недостаток не таким важным.

Все вышеизложенное доказывает, что метод на основе случайного леса лучше справляется с задачей отбора показателей в регрессионные модели, характеризующие работу нефтяных скважин.

*Статья написана при финансовой поддержке гранта «Математическое моделирование процессов нефтепереработки и нефтехимии на основе динамических моделей в терминах смесей непрерывного состава», проект № 18-47-860003 р_а.*

## Литература

1. Вирстюк А. Ю. Основные этапы создания модели эффективности работы нефтяных скважин // Наука и инновации XXI века : сб. ст. по материалам V Всерос. конф. молодых ученых : в 3 т. Сургут : ИЦ СурГУ, 2018. Т. 1. С. 19–21.

2. Афанаскин И. В., Крыганов П. В., Вольпин С. Г., Егоров А. А. Анализ добычи. Комплексирование исследований скважин и численного моделирования разработки нефтяных месторождений в рамках учебного проекта «Цифровое месторождение» / Сургут. гос. ун-т; Научно-исслед. ин-т системных исслед. Рос. Акад. Наук; Рос. Фед. Ядерный центр // Северный регион: наука, образование, культура. 2015. № 2. С. 8–18.

3. Азиз Х., Сеттари Э. Математическое моделирование пластовых систем. М. : Ин-т компьютер. Исслед., 2004. 416 с.

4. Минина И. Д. Статистика. Ч. 1. Теория статистики. Пенза: РИО ПГСХА, 2013. 225 с.

5. Feature Engineering. URL: https://nagornyy.me/courses/data-science/feature_engineering/ (дата обращения: 03.01.2020).

6. Чистяков С. П. Случайные леса: обзор // Тр. Карел. науч. центра РАН. 2013. № 1. С. 117–136.

7. Breiman L. Out-of-bag estimation // Technical report, Statistics Department University of California, Berkeley. 1996. P. 1–13. URL: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf (дата обращения: 20.03.2020).

8. Элбон К. Машинное обучение с использованием Python. Сб. рецептов. СПб. : БХВ-Петербург, 2019. 384 с.

9. Дудченко П. В. Метрики оценки классификаторов в задачах медицинской диагностики // Молодежь и современ. информ. технологии : сб. тр. XVI Междунар. науч.-практ. Конф. студентов, аспирантов и молодых ученых, 3–7 декабря 2018 г. Томск : Изд-во ТПУ, 2018. С. 164–165.

10. Ohsaki M., Wang P., Matsuda K. et al. Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification // IEEE Transactions on Knowledge and Data Engineering. 2017. No. 9. P. 1806–1819.

11. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 2017. URL: http://www.ievbras.ru/ecostat/Kiril/R/DM/DM_R.pdf (дата обращения: 09.02.2020).

12. Вирстюк А. Ю., Микшина В. С. Применение регрессионного анализа для оценки эффективности работы нефтяных скважин // Изв. Томск. политехнич. ун-та. Инжиниринг георесурсов. 2020. № 1. С.117–124.

13. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning. Data mining, inference, and prediction. 2an ed. Switzerland: Springer, 2017. 764 p.

14. Lolon E., Hamidieh K., Weijers L. Evaluating the Relationship Between Well Parameters and Production using Multivariate Statistical Models: Middle Bakken and Three Forks Case History // SPE Hydraulic Fracturing Technology Conference. The Woodlands, Texas, 2016. P. 303–331. DOI 10.2118/179171-MS.

15. Калинин А. Г. Обработка данных методами математической статистики : моногр. Чита : ЗИП СибУПК, 2015. 106 с.

16. Дронов С. В. Методы и задачи многомерной статистики : моногр. Барнаул : Изд-во АлтГУ, 2015. 275 с.

17. Халафян А. А. STATISTICA 6. Статистический анализ данных. М. : ООО «Бином – Пресс», 2007. 512 с.