Метод поиска информативных признаков для определения типа кристаллической структуры химических соединений состава AB2X2

Головин Антон, 517

Задача AB_2X_2

Набор данных полученный в ходе химических экспериментов

- 2655 соединений
- 96 физических признаков
- 17 классов соединений

Вид входных данных — соединение AB_2X_2

A, B, X — некоторые химические элементы

$$Z_i = \underbrace{f_{1,}\ldots,f_{32}}_{A},\ \underbrace{f_{33,}\ldots,f_{64}}_{B},\ \underbrace{f_{65,}\ldots,f_{96}}_{X}$$

Задача AB_2X_2.
Комбинации признаков

- Генерация нового признакового описания

  Используются обычные арифметические операции

$$Z_{new} = \{ f_4 + f_{10} * f_{65}, f_{43} - f_{91}, \ldots, f_5 - f_{91} \}$$

- Возникающие проблемы
  - Не все признаки могут между собой взаимодействовать
  - Появление сильно коррелированного признакового пространства
  - Значительное увеличение размерности признакового описания выборки
  - Метод классификации должен справляться с выборками такой величины

Случайный лес

Случайные леса(Leo Breiman,Adele Cutler)

- Классифицируют объекты
- Оценивают важность отдельных признаков
- Возвращают вероятности принадлежности элемента к классам

Разрешение проблем

- Не все признаки могут между собой взаимодействовать

  \+ Построение леса по выборке сгенерированной с учётом особенностей по взаимодействию
- Большое увеличение размерности описания выборки

  \+ Для дальнейшей генерации будут использоваться лишь наиважнейшие для классификации признаки
- Возникновение сильно коррелированного признакового описания

  \+ Conditional Feature Importance

Случайный лес

Feature importance

- Raw Feature Importance

$$Z=\begin{pmatrix} y_1 & f_{1,1} & f_{1,2} & \cdots & f_{1,j} & \cdots & p_1 & \cdots & f_{1,m} \\ y_2 & f_{2,1} & f_{2,2} & \cdots & f_{2,j} & \cdots & p_2 & \cdots & f_{1,m} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & f_{i,j} & \cdots & p_i & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_n & f_{n,1} & f_{n,2} & \cdots & f_{n,j} & \cdots & p_n & \cdots & f_{n,m} \end{pmatrix} \longrightarrow Z^{new}=\begin{pmatrix} y_1 & f_{1,1} & f_{1,2} & \cdots & f_{g_j(1),j} & \cdots & p_1 & \cdots & f_{1,m} \\ y_2 & f_{2,1} & f_{2,2} & \cdots & f_{g_j(2),j} & \cdots & p_2 & \cdots & f_{1,m} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & f_{g_j(i),j} & \cdots & p_i & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_n & f_{n,1} & f_{n,2} & \cdots & f_{g_j(n),j} & \cdots & p_n & \cdots & f_{n,m} \end{pmatrix}$$

$$RawImp(j)=\sum_{i=1}^{n} I(y_i=Tree(z_i))-\sum_{i=1}^{n} I(y_i=Tree(z_i^{g(j)}))$$

- Conditional Feature Importance

$$Z=\begin{pmatrix} y_1 & f_{1,1} & f_{1,2} & \cdots & f_{1,j} & \cdots & p_1 & \cdots & f_{1,m} \\ y_2 & f_{2,1} & f_{2,2} & \cdots & f_{2,j} & \cdots & p_2 & \cdots & f_{1,m} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & f_{i,j} & \cdots & p_i & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_n & f_{n,1} & f_{n,2} & \cdots & f_{n,j} & \cdots & p_n & \cdots & f_{n,m} \end{pmatrix} \longrightarrow Z^{new}=\begin{pmatrix} y_1 & f_{1,1} & f_{1,2} & \cdots & f_{(g_j(1),P=a),j} & \cdots & p_1=a & \cdots & f_{1,m} \\ y_3 & f_{3,1} & f_{3,2} & \cdots & f_{(g_j(3),P=a),j} & \cdots & p_3=a & \cdots & f_{3,m} \\ y_{25} & f_{25,1} & f_{25,2} & \cdots & f_{(g_j(25),P=a),j} & \cdots & p_{25}=a & \cdots & f_{25,m} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & f_{(g_j(i),P=b),j} & \cdots & p_i=b & \cdots & \vdots \\ y_{55} & f_{55,1} & f_{55,2} & \cdots & f_{(g_j(55),P=b),j} & \cdots & p_{55}=b & \cdots & f_{55,m} \\ y_n & f_{n,1} & f_{n,2} & \cdots & f_{(g_j(n),P=b),j} & \cdots & p_n=b & \cdots & f_{n,m} \end{pmatrix}$$

$$CondImp(j)=\sum_{i=1}^{n} I(y_i=Tree(z_i))-\sum_{i=1}^{n} I(y_i=Tree(z_i^{g(j),P}))$$

Модельный пример

- Выборка
  - 200 элементов
  - 10 признаков
  - 2 класса

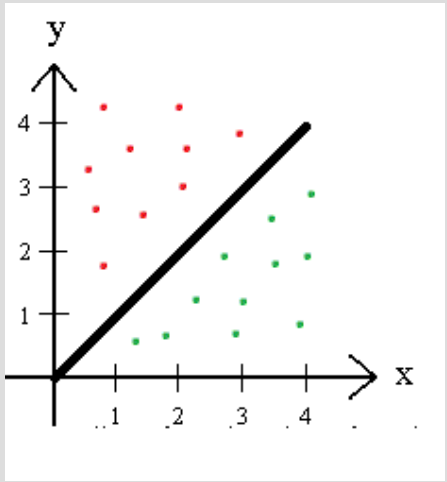
Все признаки сгенерированы из равномерного распределения от 0 до 5.

Метка класса

x – первый признак
y – второй признак
если x-y > 0 метка 1
иначе метка 0

Добавляем сильно коррелирующий признак x + 0.01*y

Другие признаки - шум



- Raw Feature Importance

Feature x:	1,86
Feature y:	2,35
Feature x+0.01*y:	2,18

- Conditional Feature Importance

Feature x:	0,18
Feature y:	0,22
Feature x+0.01*y:	0,02

glmnet

Решается задача линейной регрессии.

$$\min_{(\beta_0,\beta)\in R^{p+1}}\left[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda P_\alpha(\beta)\right]$$

$$P_\alpha(\beta)=(1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2+\alpha\|\beta\|_{l_1}=\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|\right]$$

glment
Оценка важности признаков

1) Из исходной генерируется новая обучающая выборка и OOB тестовая выборка.
2) Обучение → вектора весов для каждого значения коэффициента регуляризации
3) Сопоставление признакам значения их ранга.
4) Вычисление среднего ранга признаков по всем выборкам

$$A_1 = \begin{bmatrix} & \beta_{\lambda_1} & \beta_{\lambda_2} & \beta_{\lambda_3} \\ \beta_1 & 0 & 0 & 0 \\ \beta_2 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \beta_i & 0 & 0.237833 & -0.5637833 \\ \cdots & \cdots & \cdots & \cdots \\ \beta_j & 0 & 0 & -1.42543 \\ \cdots & \cdots & \cdots & \cdots \\ \beta_m & 0 & 0 & 0 \end{bmatrix} \rightarrow FeatureRank(A_1) = \{\beta_i = 1, \beta_j = 2, \ldots\}$$

$$\hat{FeatureRank} = \sum_{i=1}^{N} \frac{FeatureRank(A_i)}{N}$$

Итерационная схема генерирования признаков

1)Обучение по исходной выборке. Отбор N наиболее информативных признаков

- Случайный лес
- glmnet

2)Генерация новой выборки с использованием только N наиболее информативных признаков

3) Оценка обобщающей способности, ошибка OOB

Сравнение glment и RForest

Выборка тестовых данных:
1000 объектов, 100 признаков, 2 класса
Rforest обучение 500 деревьев 37 секунд
OOB error: 23.3
Glment обучение 9 секунд
OOB error: 17.7

Спасибо за внимание!!!

Задача AB_2X_2
Результаты

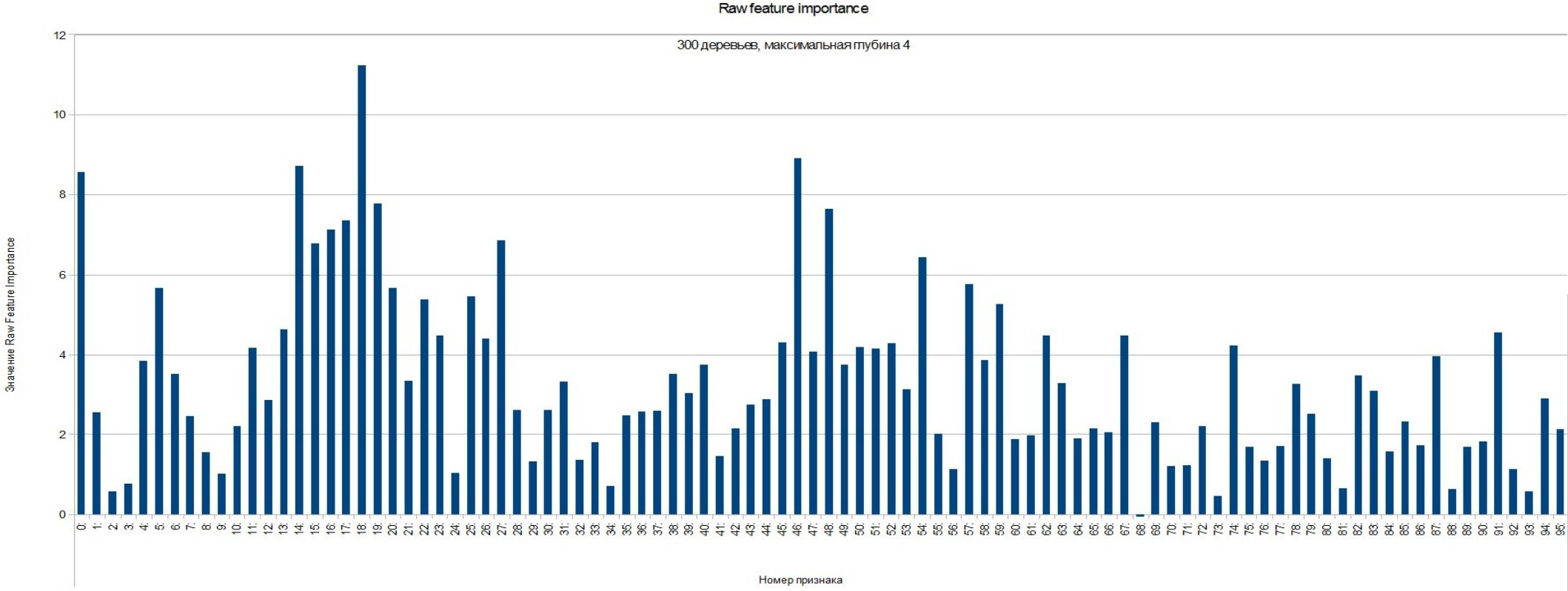
Задача AB_2X_2 на 17 классов для 2655 элементов с 96 признаками.

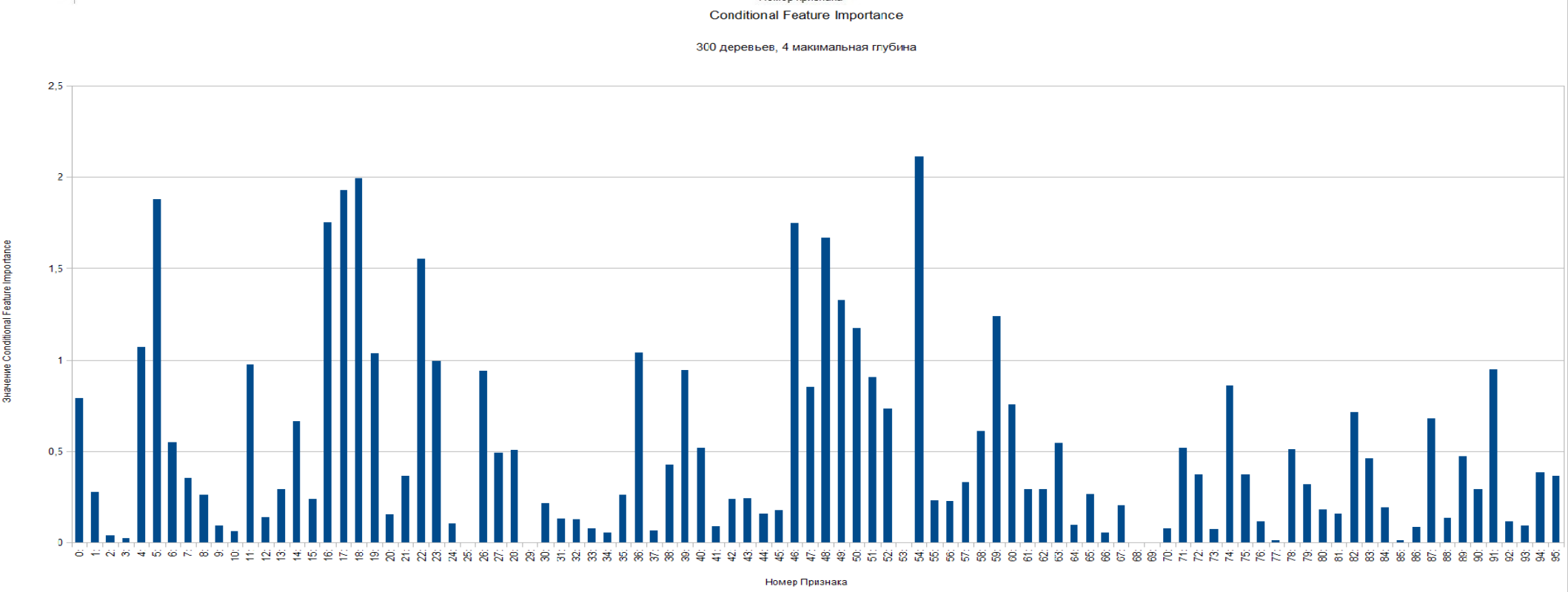
- 10 наиболее информативных признаков по критерию Raw Feature Importance:

0, 13, 14, 15, 16, 17, 18, 19, 46, 48
Ошибка классификации: 0,27

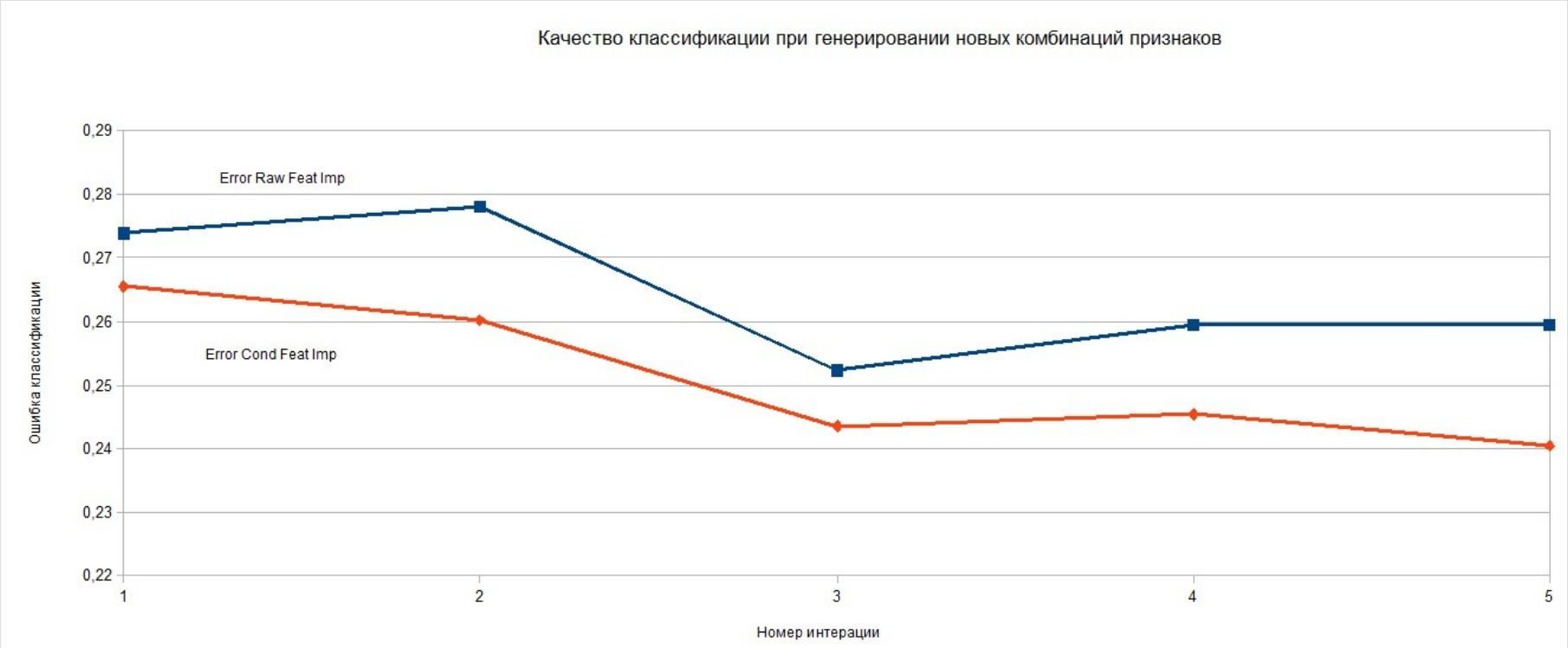
- 10 наиболее информативных признаков по критерию Conditional Feature Importance:

5, 16, 17, 18, 22, 46, 48, 49, 54, 60
Ошибка классификации: 0,24


Raw feature importance
300 деревьев, максимальная глубина 4
Значение Raw Feature Importance
Номер признака


Conditional Feature Importance
300 деревьев, 4 макимальная ггубина
Значение Conditional Feature Importance
Номер Признака

Сравнение качества классификации при генерации признаков



Сравнение скорости работы