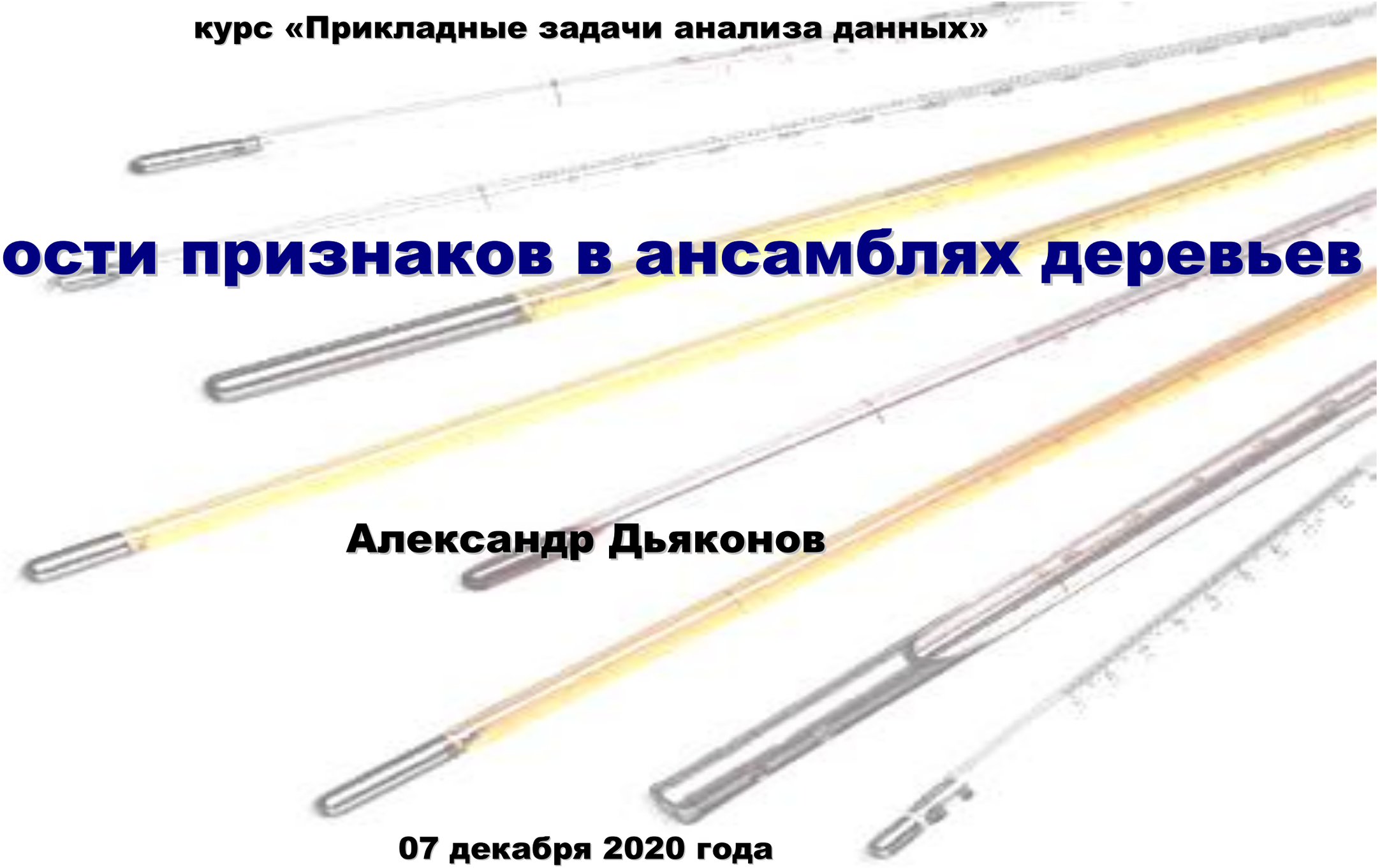
курс «Прикладные задачи анализа данных»



Важности признаков в ансамблях деревьев

Александр Дьяконов

07 декабря 2020 года

План

Проблема формализации важности признаков

Примеры использования важности признаков

Важность по неоднородности (impurity-based importance)

Перестановочная важность PFI (Permutation Feature Importance)

Эксперименты по оцениванию важности

Boruta (идея)

ACE (Artificial Contrasts with Ensembles)

Важность признаков

Важность признаков – числовые оценки, насколько каждый признак важен для решения поставленной задачи

- **влияет на полученный результат**

т.е. на значение модели $\Rightarrow$ зависимость от модели

- **обеспечивает качество решения задачи**

т.е. входит в «невидимые паттерны»

Напоминание…

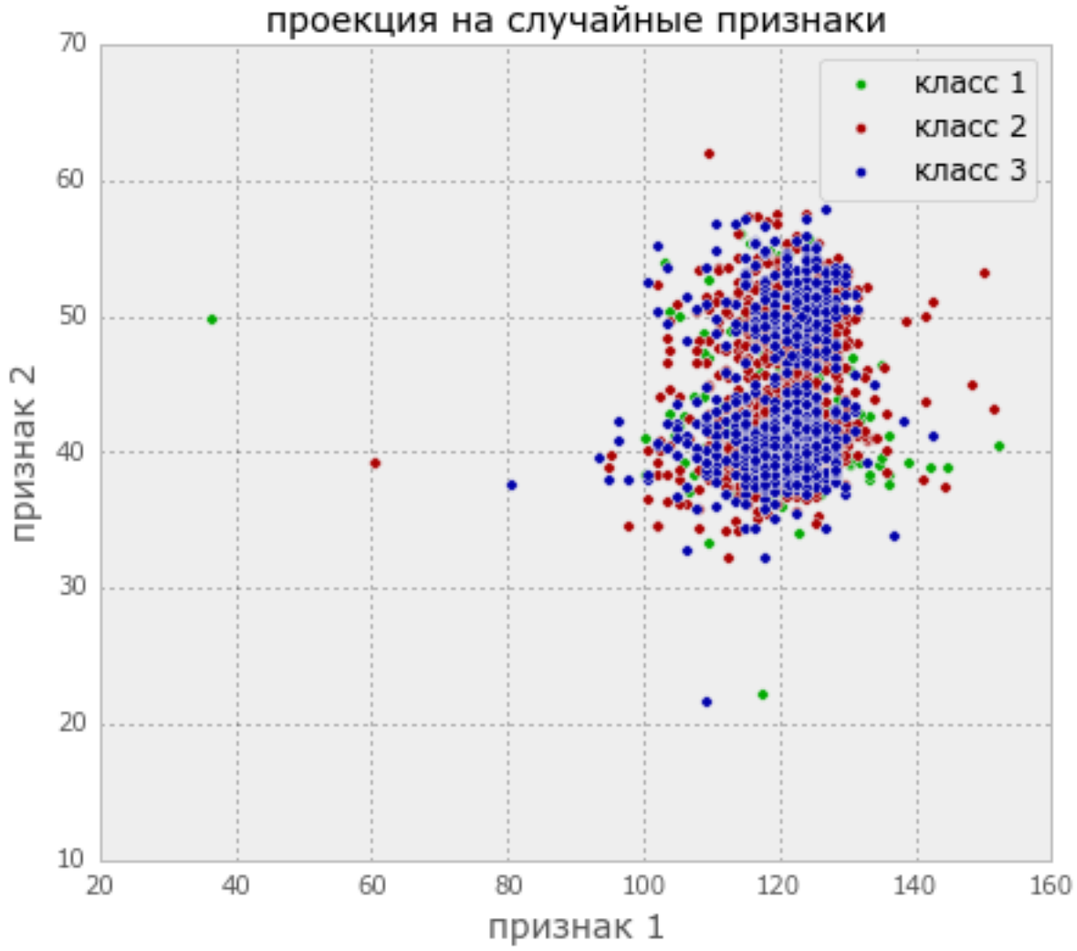
селекция признаков производится с помощью

- **фильтров**
- **обёрток**
- **встроенных методов**

на всех группах возможна оценка качества

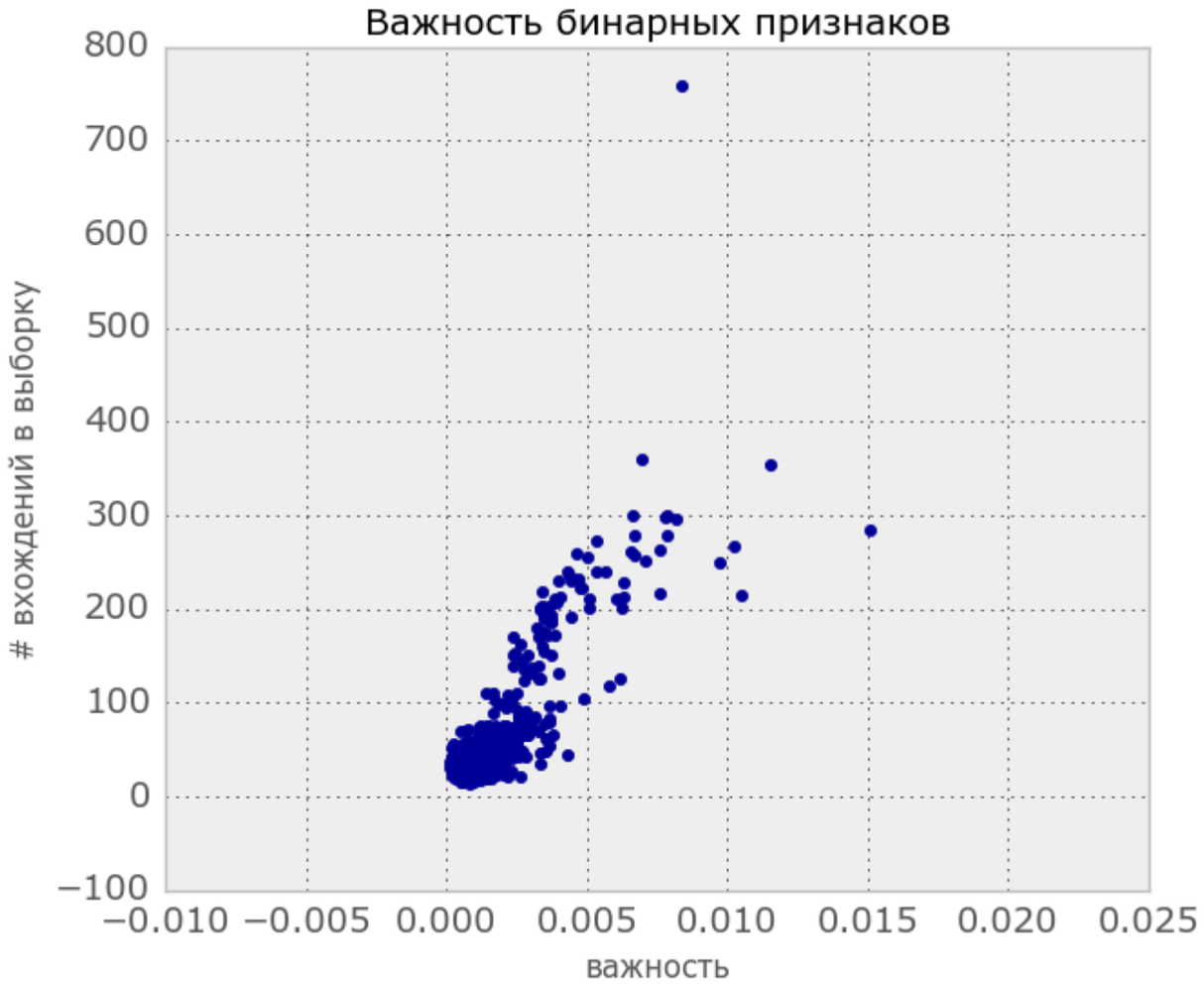
дальше только про методы оценки важности при использовании ансамблей деревьев

Пример использования важности признаков: поиск хороших признаков



Быстро понять от чего зависит целевой признак

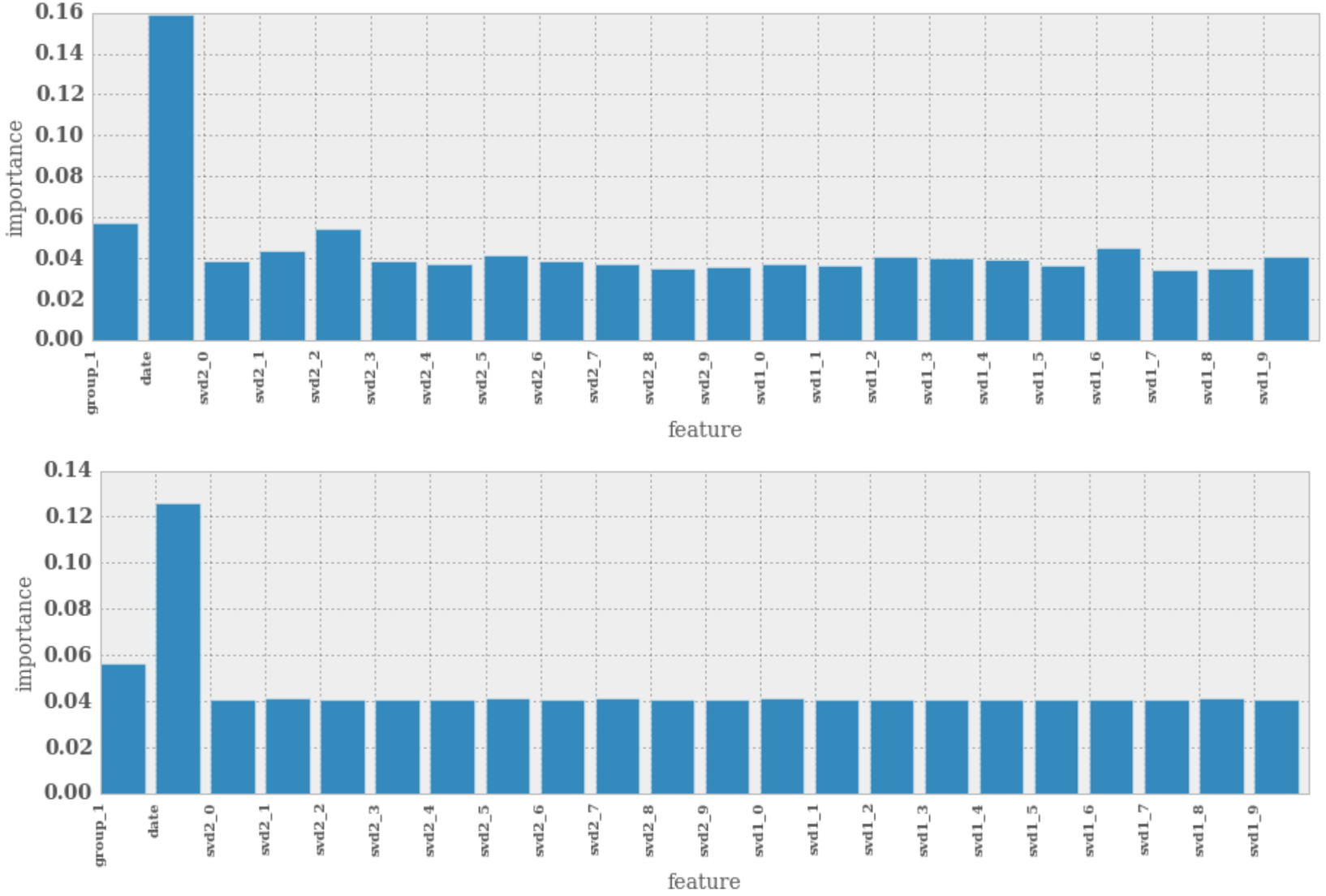
Пример использования важности признаков: EDA



По вертикали – число ненулевых значений признака

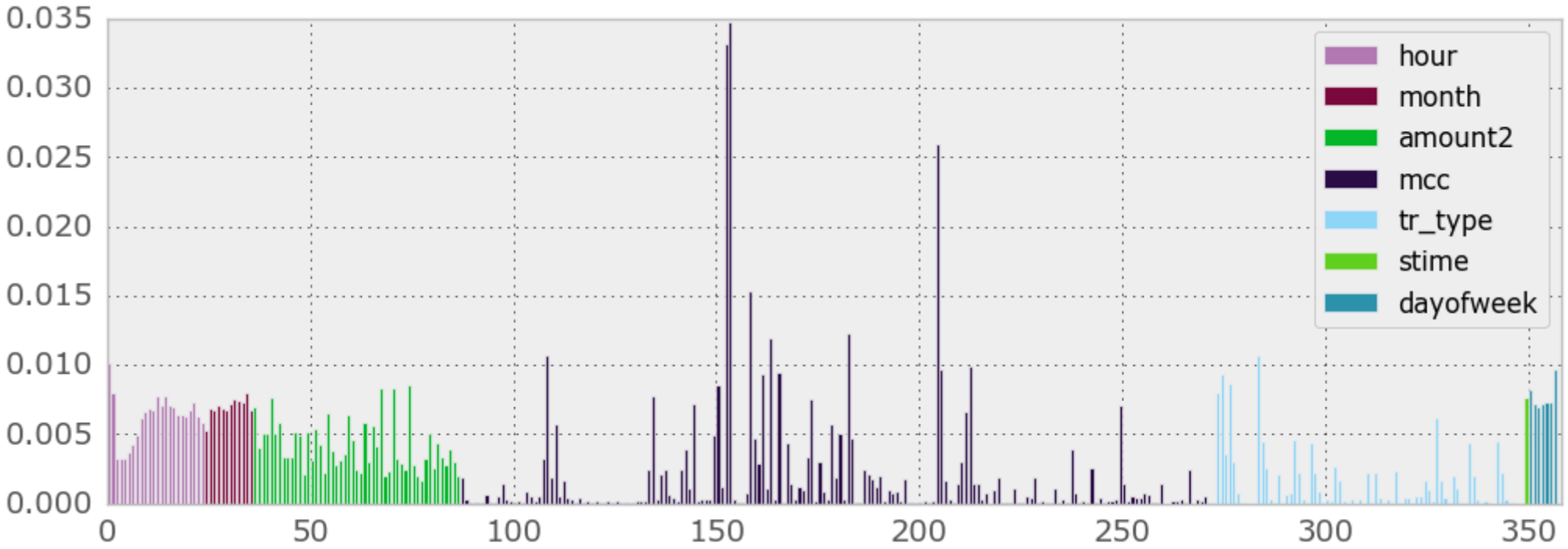
Видна группа очень неплохих признаков;)

Пример использования важности признаков: проверка гипотез



Как отличаются важности случайных и SVD-признаков

Пример использования важности признаков: задача СберБанка

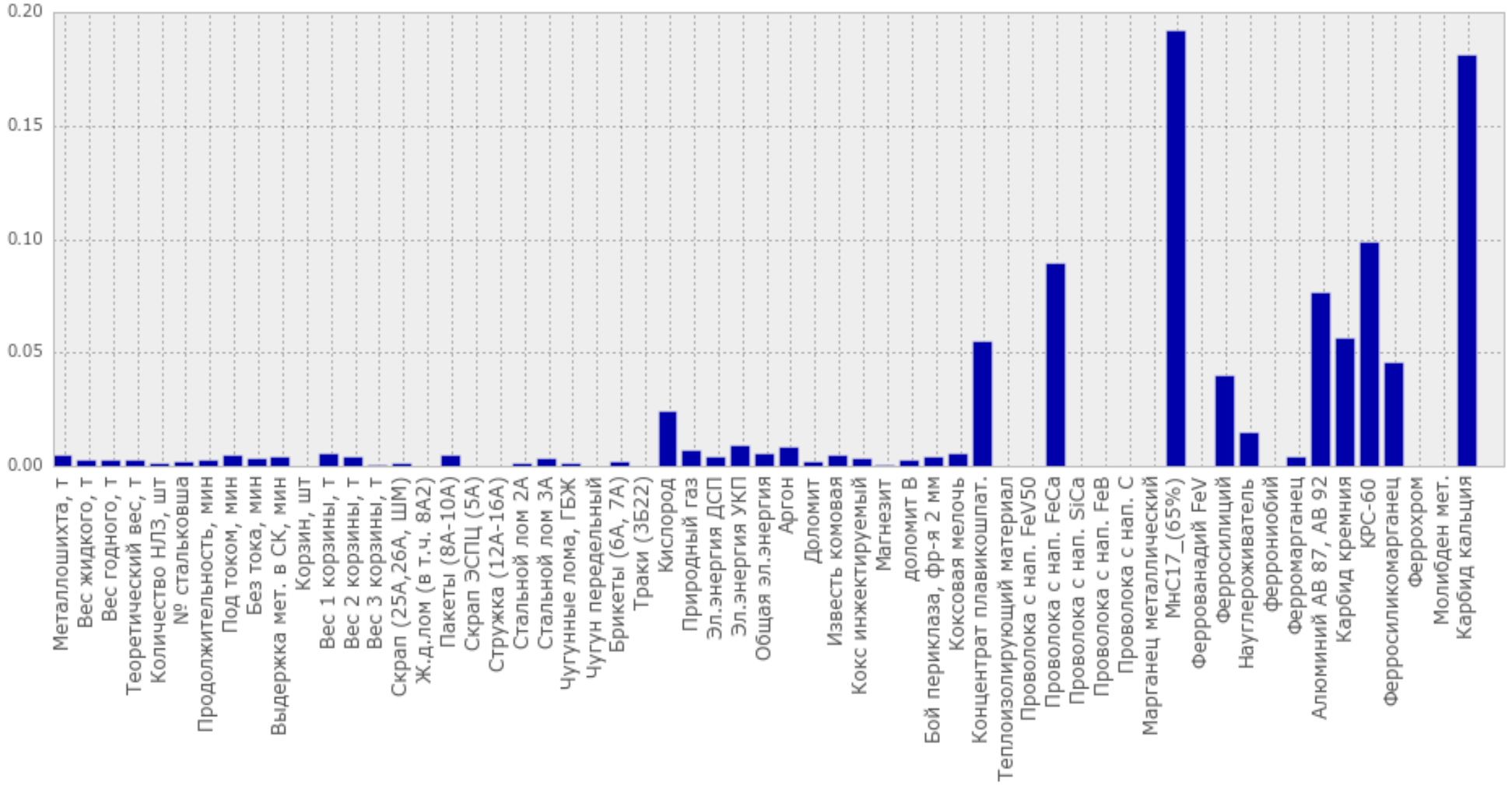


`rf.feature_importances_`

```
rf = RandomForestClassifier(n_estimators=1000, max_features=30, n_jobs=-1)
rf.fit(X, y)
plt.bar(np.arange(len(rf.feature_importances_)),
        rf.feature_importances_, color='black')
```

Можно сразу увидеть важные признаки и целые группы…

Пример использования важности признаков: Металлургия



сразу понятно, от чего зависит целевой признак

Как вычислить важность?

Плохой метод – чем чаще выбирался признак, тем лучше.

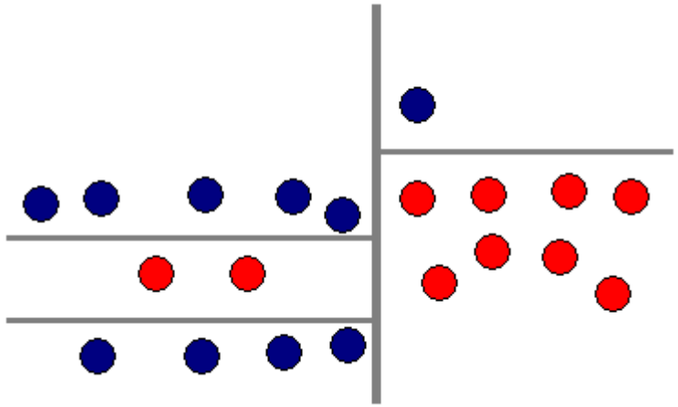
Почему?

Как использовать важность?

Не увлекаться выбрасыванием неважных признаков

Почему?

Как вычислить важность?



По хорошим признакам меньше всего расщеплений…

Как использовать важность?

не увлекаться выбрасыванием неважных признаков:

- **оценка качества признаков не всегда адекватная**
- **если много хороших коррелированных признаков, то их важность будет маленькая**
- **не рекомендуют оценивать важность и решать одним и тем же алгоритмом**

Важность по неоднородности (impurity-based importance)

Уменьшение неоднородности по признаку при расщеплении:

$$Q(R,\theta) = H(R) - \frac{|R_{\text{left}}|}{|R|} H(R_{\text{left}}) - \frac{|R_{\text{right}}|}{|R|} H(R_{\text{right}})$$

для каждого признака – сумма УН по всем расщеплениям, в которых он участвовал (во всех деревьях)

иногда + нормировка

формула обобщается и на случай весовой выборки

см. код – может использоваться

$$\frac{|R|}{m}\left(H(R) - \frac{|R_{\text{left}}|}{|R|} H(R_{\text{left}}) - \frac{|R_{\text{right}}|}{|R|} H(R_{\text{right}}) \right)$$

Breiman Random Forests // Machine Learning, 45(1), 5-32, 2001

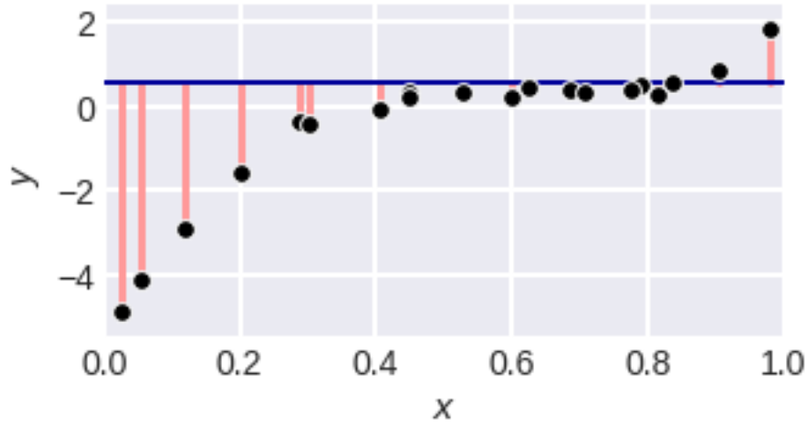
Важность по неоднородности (impurity-based importance)

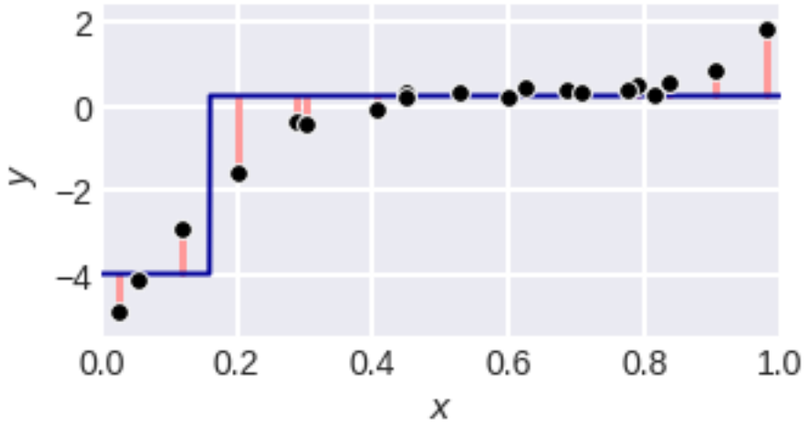
другие названия:

Gini importance (в sklearn)

IncNodePurity (в R)

MDI – mean decrease impurity (как раз, когда усредняют по деревьям)





+ автоматически вычисляется при построении деревьев (реализован в sklearn)

– годится только для деревьев / их ансамблей

– смещённость в сторону признаков с большим числом значений, также в случае разномасштабности признаков

Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T. Bias in random forest variable importance measures // BMC Bioinformatics. 2007. 8(1), 25.

Оценка важности: другая идея

признак важный, если его «модификация» снижает качество

Какая модификация?

Оценка важности: другая идея

признак важный, если его «модификация» снижает качество

Какая модификация?

перестановка значений (Permutation Feature Importance)

рассмотрим дальше

+ не надо переобучать модель – смотрим как меняется качество на отложенной выборке

удаление признака (Drop-column importance)

– надо переобучить модель

+ более однозначный результат (см. дальше)

иногда использую его:

Parr. T., Turgutlu K., Csiszar C., Howard J. Beware Default Random Forest Importances // https://explained.ai/rf-importance/

Перестановочная важность PFI (Permutation Feature Importance)

Идея: признак важный, если перестановка его значений снижает качество

Есть реализация в scikit-learn

- **изменение качества на обучении**
- **изменение качества на отложенном контроле**
- **изменение качества на любой схеме валидации**

+ перестановка не меняет распределение значений

+ можно применять на любых алгоритмах

+ соответствует интуиции

+ самый надёжный метод

– очень медленный

– в чём ещё подвохи?

+ в бутстрепе можно использовать OOB-контроль!

Перестановочная важность PFI (Permutation Feature Importance)

PFI на OOB

1. **Вычисляем качество Q на OOB**
2. **Для i-го признака – делаем случайную перестановку значений, вычисляем качество Q_i на OOB**
3. **Информативность i-го признака =** $\max(Q - Q_i, 0)$

Запомните приём!

Вместо качества можно использовать что-то другое…

~ доля верно классифицирующих деревьев

другие названия: %IncMSE (в R)

Замечания: сравнивать скорость вычисления важностей в R и scikit-learn некорректно

Перестановочная важность PFI (Permutation Feature Importance)

Подвох 1 – несколько сильно коррелированных признаков

перестановка значений одного из них может слабо влиять на качество решения

вариант решения – кластеризация признаков (по функции схожести «корреляции») и формирование признакового пространства из представителей кластеров

https://scikit-learn.org/stable/

Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A. Conditional Variable Importance for Random Forests. 2008. BMC Bioinformatics, 9(1), 307.

Подвох 2 – оценка зависит от перестановки

На практике делают несколько (ещё медленнее)

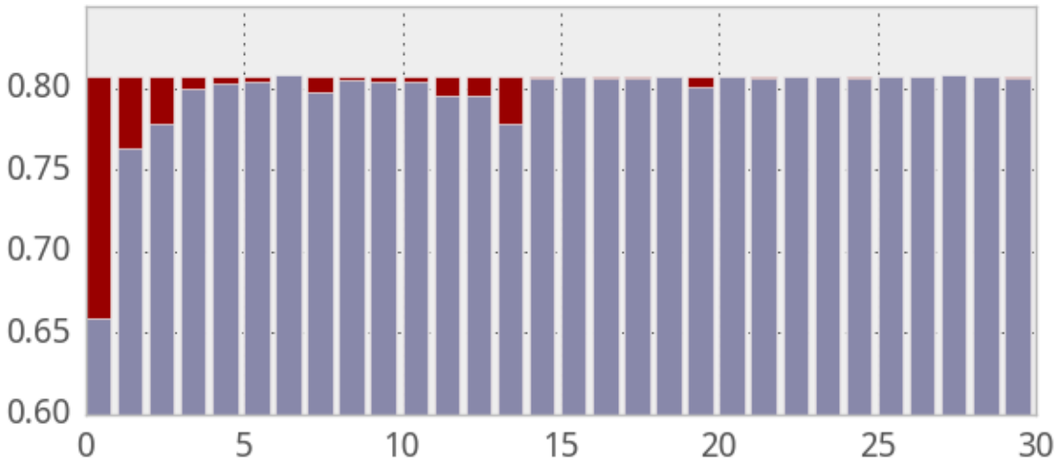
Подвох 3 – метод универсальный, но зависит от конкретной модели и критерия качества

Перестановочная важность PFI (Permutation Feature Importance)

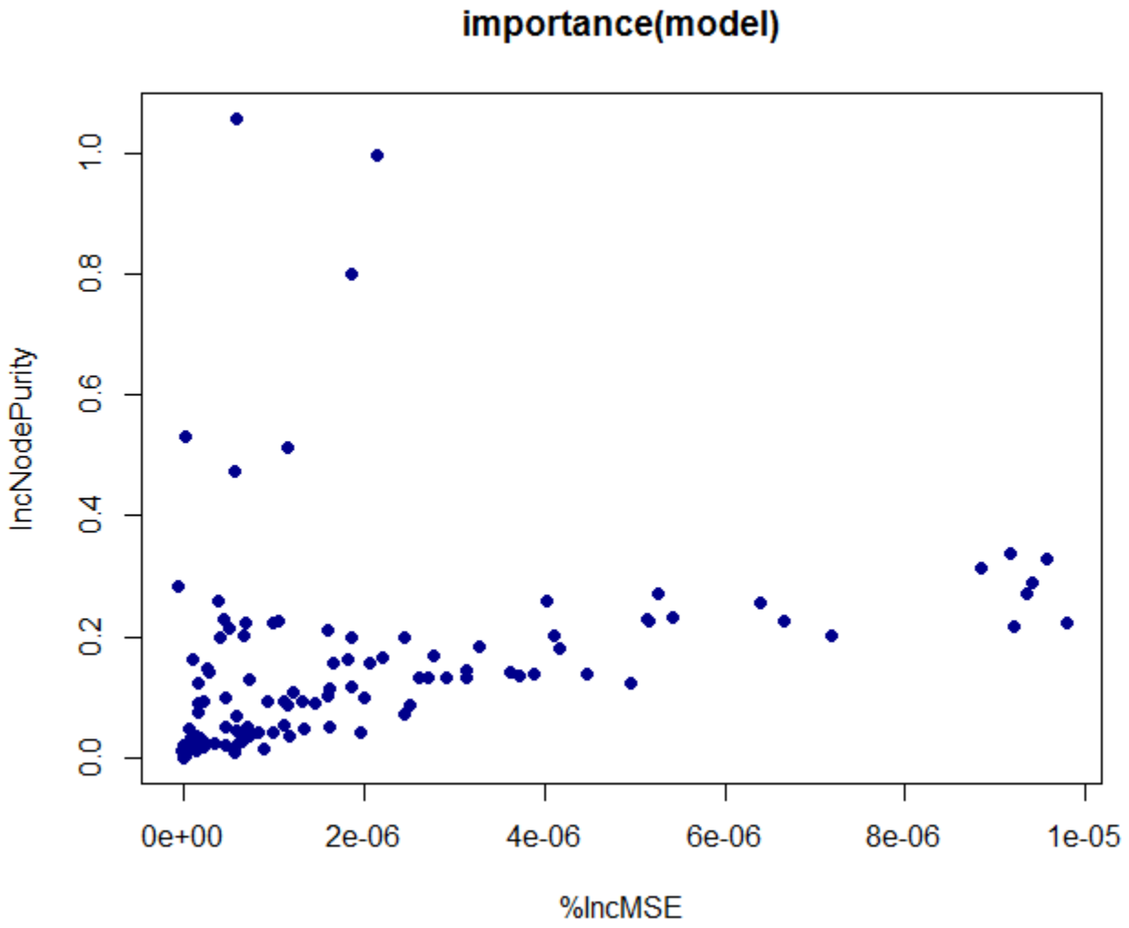
возможны простые реализации

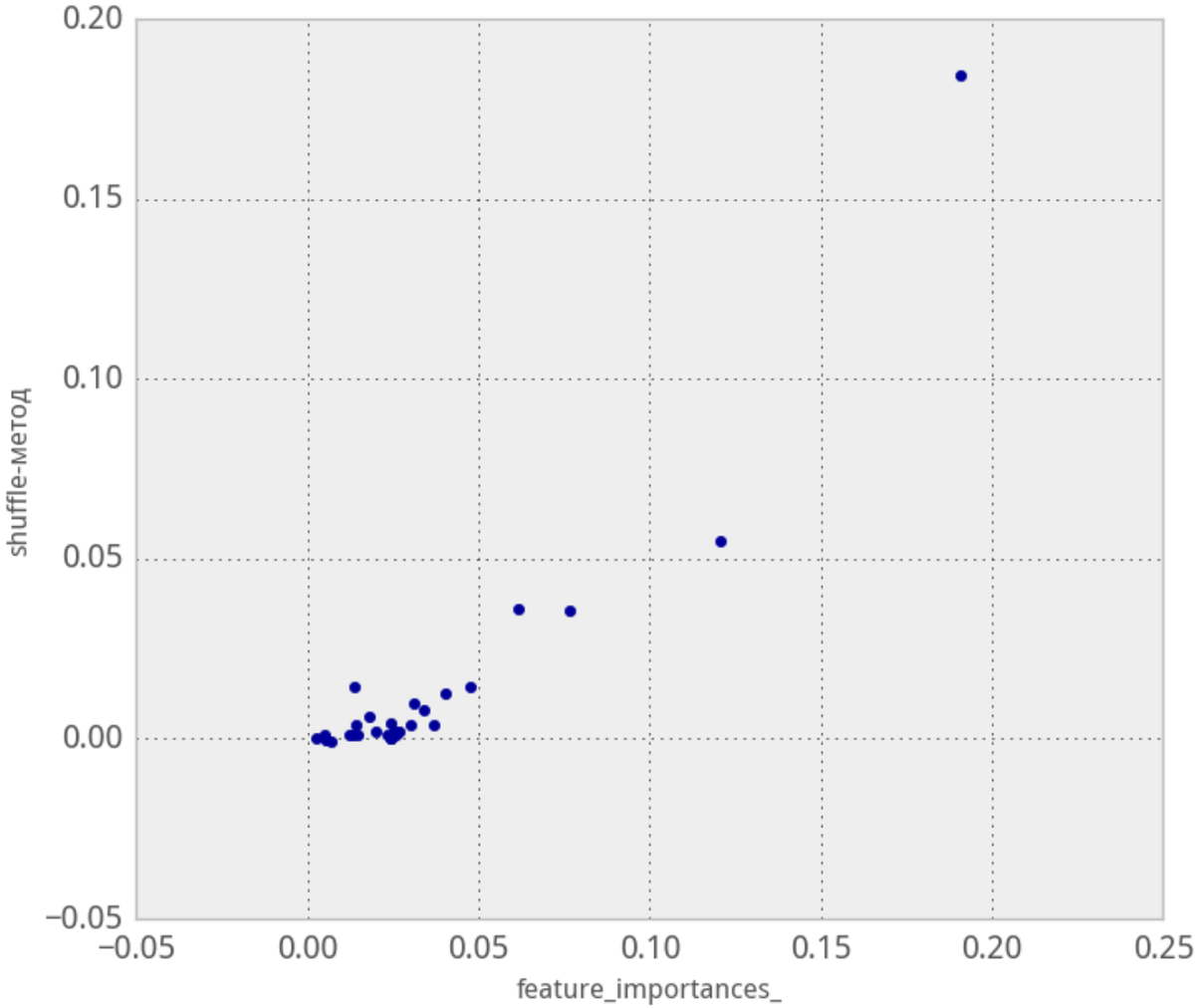
```
e = [] # качество классификации
a = rf.predict(X2)
q = roc_auc_score(y2, a) # базовое качество
классификации
for t in range(X2.shape[1]):
    Xt = X2.copy()
    np.random.shuffle(Xt[:, t]) # перемешиваем
    at = rf.predict(Xt)
    e.append(roc_auc_score(y2, at))

e = np.array(e)
plt.bar(np.arange(len(e)), e*0 + q, color =
'#990000')
plt.bar(np.arange(len(e)), e, color = '#8888AA')
```

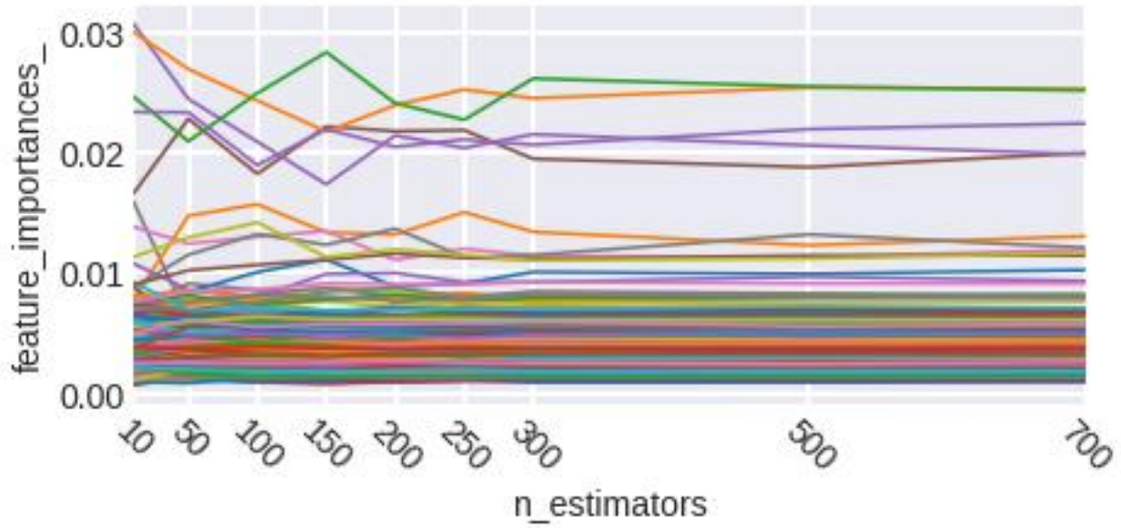


Сравнение разных важностей: задача скоринга / ?





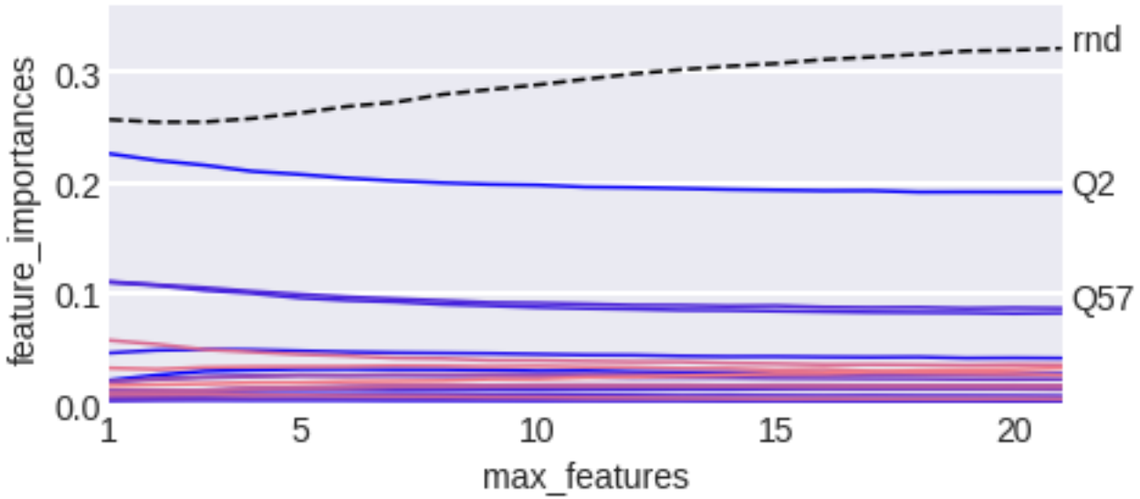
Эксперименты по оцениванию важности



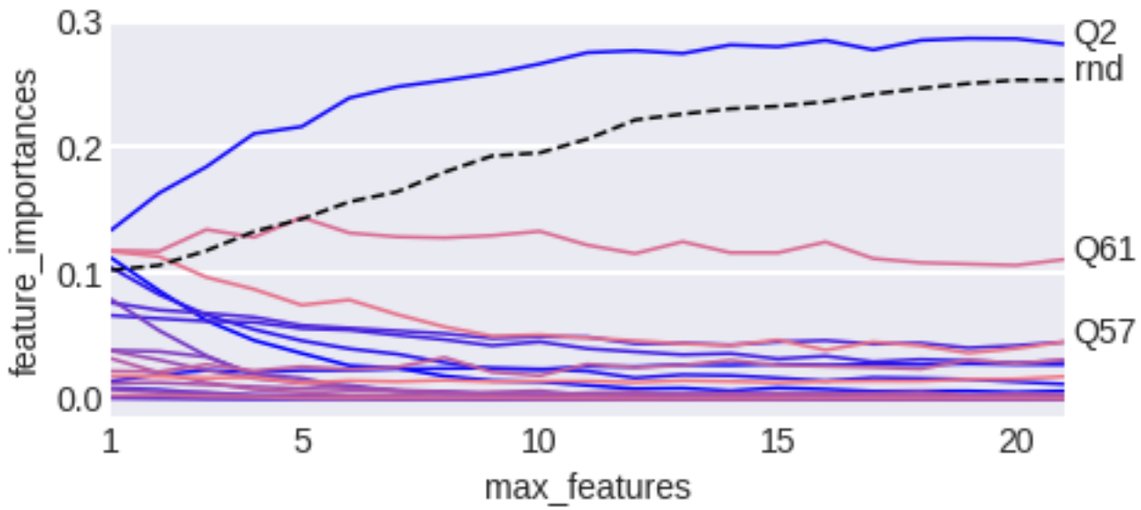
При увеличении числа деревьев есть сходимость!

Эксперименты по оцениванию важности

MDI



PFI-train



использование обучения для оценки важности в PFI

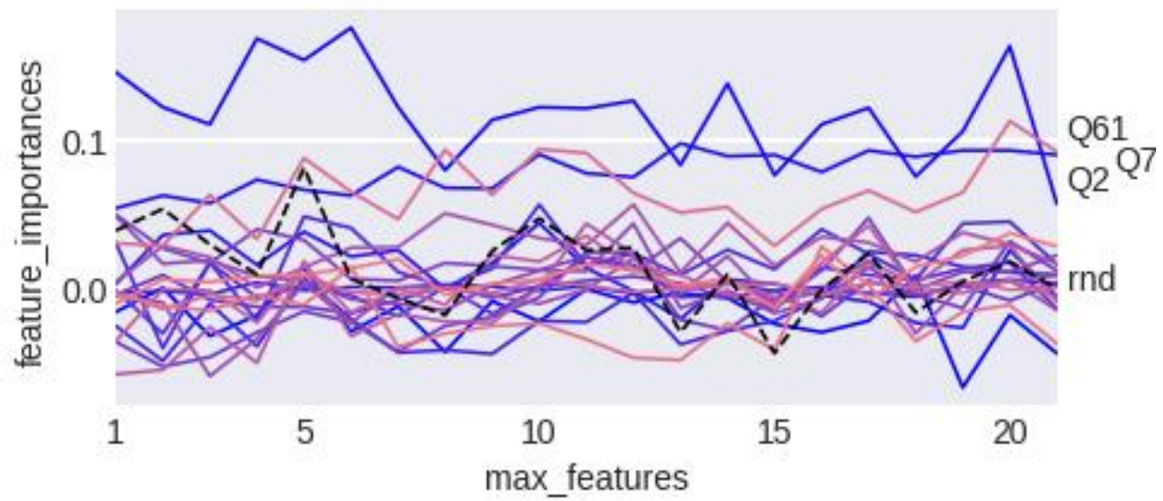
Случайный признак «rnd» U[0,1] – много уникальных значений

Эффект важности случайного признака известен (но тут он совсем большой + возможен при неправильном использовании PFI)

Parr. T., Turgutlu K., Csiszar C., Howard J. Beware Default Random Forest Importances // https://explained.ai/rf-importance/

Эксперименты по оцениванию важности

PFI-holdout



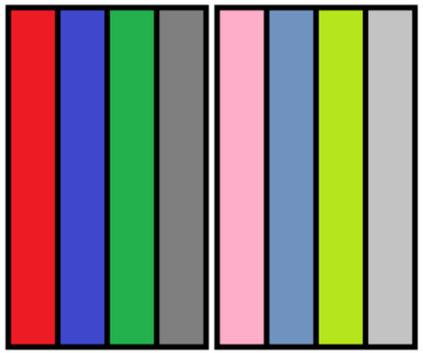
10×20%holdout×2permutations

PFI-holdout-разные критерии расщепления и разные модели (усреднение по набору параметров)

Boruta (идея)

1. **Добавить к исходным признакам их перемешанные (shuffle) копии**
   (shadow features / признаки-призраки)

признаки **перемешали**



2. **Запустить RF – вычислить Z-меру, MSZA = max(Z-score) на перемешанных признаках**

Здесь важность – потеря точности классификации, вычисляется для каждого дерева, из всех содержащих рассматриваемый признак

Приём – shuffle

Boruta (идея)

3. Запустить RF на исходных данных

Если Z-score << MSZA, то признак плохой
Если Z-score >> MSZA, то признак хороший

Можно удалить плохие признаки и повторить процедуру

Что такое Z-score

$$z = \frac{x - \mu}{\sigma}$$

здесь ~ rf_importance / стандартное отклонение
(средняя важность и отклонение считается по всем деревьям)

ACE (Artificial Contrasts with Ensembles)

Аналогично, но удаляются хорошие признаки!

Советы

- оценивать важность не обязательно с помощью лучшей модели

- перестановочная важность самая естественная

но есть нюансы (коррелированность признаков, стабильность и надёжность оценки и т.п.)

- есть другие подходы и удобные библиотеки (SHAP)

в лекции про интерпретацию моделей

Литература

Обзорная статья про интерпретацию
https://dyakonov.org/2018/08/28/интерпретации-чёрных-ящиков/

Miron B. Kursa, Witold R. Rudnicki Feature Selection with the Boruta Package // JSS 2010, DOI 10.18637/jss.v036.i11, https://www.jstatsoft.org/article/view/v036i11