Случайный лес и оценка значимости признаков

Юлия Пономарева
Data Scientist

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

Цели занятия

1. Изучим разложение ошибки на смещение и разброс
2. Реализуем свой бэггинг
3. Узнаем устройство модели случайный лес
4. Подберем оптимальные гиперпараметры для случайного леса
5. Разберемся с OOB-ошибкой
6. Вспомним методы оценки значимости признаков

План занятия

1. Разложение ошибки на смещение и разброс
2. Бэггинг
3. Случайный лес
4. OOB-ошибка
5. Оценка важности признаков
6. Итоги занятия

МФТИ

Практика
(Ансамбли)

Оценка значимости признаков

Методы отбора признаков

Фильтры (одномерный отбор)
основаны на некоторых показателях, которые не зависит от метода классификации (коэффициент корреляции, взаимная информация, F-тест, Хи-квадрат)

Обертки
опираются на информацию о метрике качества, полученную от моделей ML (последовательный отбор и последовательное исключение признаков и др.)

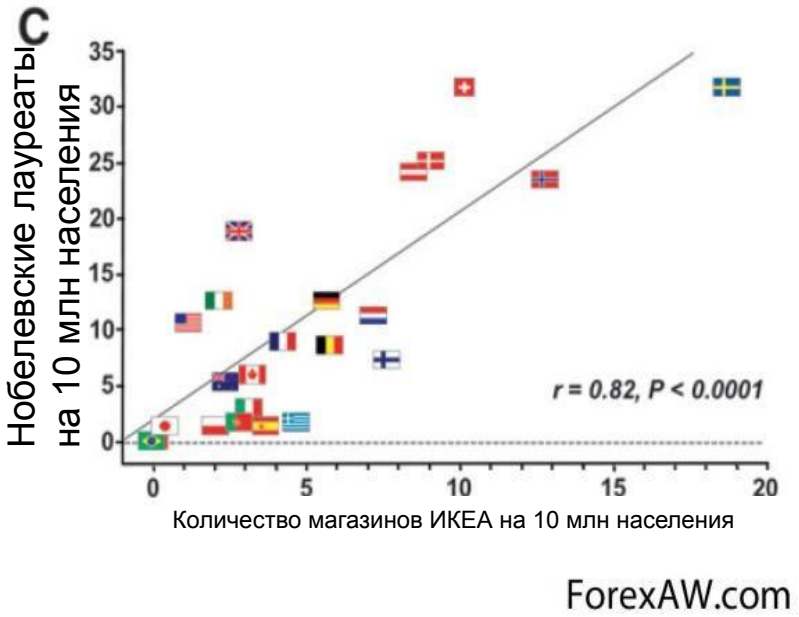
Встроенные в алгоритмы
выполняют отбор признаков во время процедуры обучения классификатора, и именно они явно оптимизируют набор используемых признаков для достижения лучшей точности (регрессия с L1-регуляризация, Random Forest, SHAP, перемешивания и др.)

Фильтры

Корреляция

Коэффициент корреляции Пирсона

$$R_{k_i,p} = \frac{\sum_{i=1}^{n}(k_i - \hat{k})\cdot(p_i - \hat{p})}{\sqrt{\sum_{i=1}^{n}(k_i - \hat{k})^2 \cdot \sum_{i=1}^{n}(p_i - \hat{p})^2}}$$



Взаимная информация (Mutual Information)

Чем выше значение MI, тем сильнее связь между этой переменной и таргетом, что говорит о том, что мы должны поместить эту переменную в набор данных для обучения

$$Entropy = -\sum p(X) \log p(X)$$

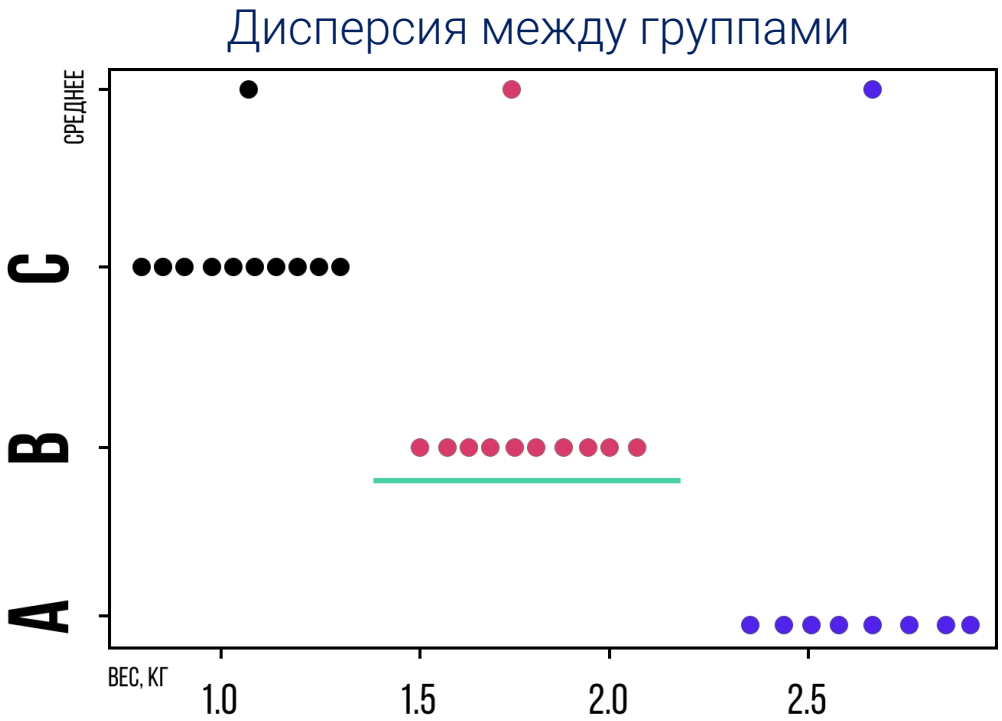
Зависимость между полом и пользованием страховыми услугами

Пол	Пользуетесь ли Вы услугами страхования жизни?	
	Да	Нет
Мужской	39%	54%
Женский	61%	46%
Итого по столбцу	100%	100%

F-тест (критерий Фишера)

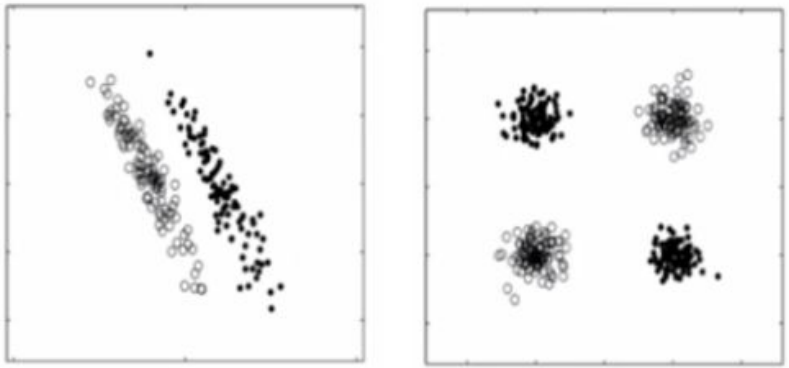
Чем больше F, тем проще различить выборки

$$\frac{\text{Дисперсия между группами}}{\text{Дисперсия внутри группы}}$$



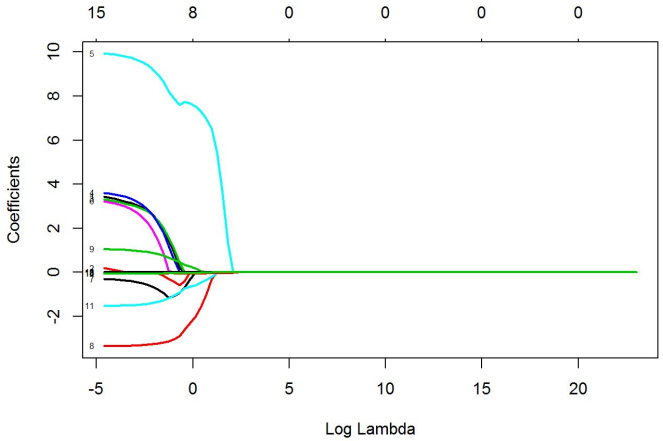
Одномерный отбор

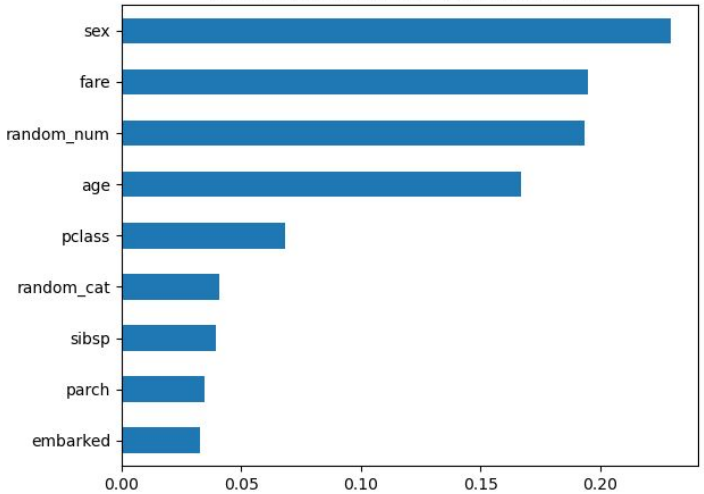
У одномерного отбора признаков есть проблема - они не учитывают взаимосвязь признаков, зависимость целевой переменной от сложной комбинации признаков.



Встроенные в модель

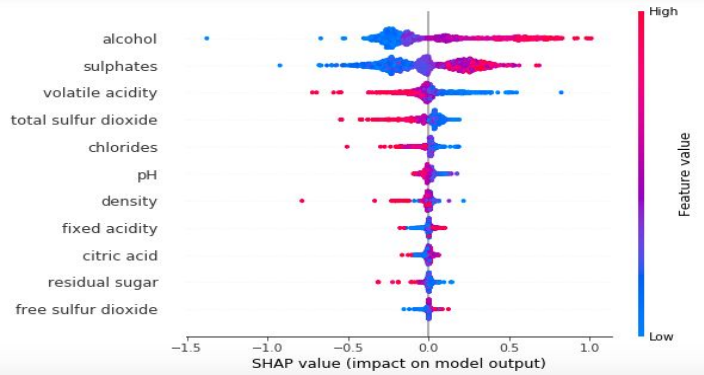
Методы встроенные в алгоритмы





Методы встроенные в алгоритмы (отдельная библиотека - SHAP*)

SHAP – значения показывают, насколько данный конкретный признак изменил наше предсказание (по сравнению с тем, как мы сделали бы это предсказание при некотором базовом значении этого признака)

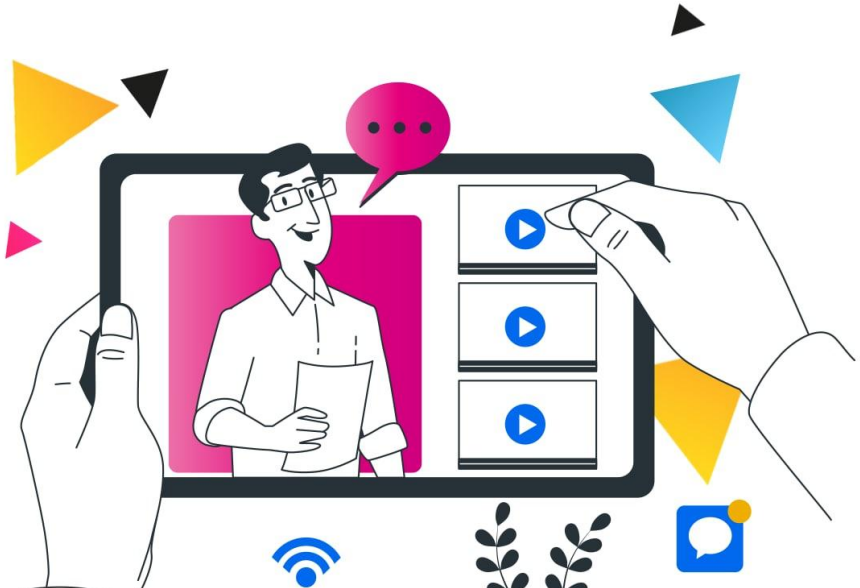


МФТИ

Практика была

https://colab.research.google.com/drive/1aUW_cYoSkhHNW177nTaL82IuQt892Eha

Ваши вопросы?



Итоги занятия

Итоги занятия

1. Изучили разложение ошибки на смещение и разброс
2. Реализовали свой бэггинг
3. Узнали устройство модели случайный лес
4. Подобрали оптимальные гиперпараметры для случайного леса
5. Разобрались с OOB-ошибкой
6. Вспомнили методы оценки значимости признаков

Дополнительные материалы

1. Ансамбли в машинном обучении - https://dyakonov.org/2019/04/19/%D0%B0%D0%BD%D1%81%D0%B0%D0%BC%D0%B1%D0%BB%D0%B8-%D0%B2-%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%BC-%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B8/
2. Ансамблевые методы: бэггинг, бустинг и стекинг - https://neurohive.io/ru/osnovy-data-science/ansamblevye-metody-begging-busting-i-steking/
3. Бэггинг и бутстрап + композиции в целом - https://habr.com/ru/company/ods/blog/324402/
4. Бэггинг и случайный лес - https://youtu.be/rawnIo_XtYY

МФТИ

Пожалуйста, оставьте свой отзыв о семинаре

До встречи!