26.10.2023

Методы оценки эффективности Black Box моделей

Начальник Центра валидации моделей
и рейтинговых систем
Алексей Чебыкин

настоящие возможности

росбанк 30 лет

01

Black box.
Что это такое?

Поговорим об основных понятиях

Black Box или Черный ящик

В общем случае – это некая модель, о внутренностях которой мы:

- Либо не имеем представления;
- Либо в которые не хотим или не можем погрузиться;

Таким образом, под это определение подходят как настоящие black box модели (например, внешние скоринги), так и сложные слабоинтерпретируемые модели (градиентные бустинги, нейронные сети и т.д.).

Для наших целей воспользуемся следующей классификацией black box моделей

Статичные – модели, существующие as-is, то есть на работу которых мы не можем воздействовать - только подавать на вход данные.

Управляемые – модели, на работу которых мы можем воздействовать, подавая набор управляющих параметров.

Обучаемые – модели, на работу которых мы можем воздействовать, переобучая модель и получая новую, чем-то похожую.

02

Метрики. Базовая оценка эффективности.

Метрики

Для каждой задачи есть свои метрики (и мы их прекрасно знаем), которые мы можем использовать для точечной оценки.

Бинарная классификация.

Gini, ROC_AUC, Precision, Recall, Accuracy, F-score….

Регрессия.

MAE, MAPE, R_square, RMSE…..

И так далее…

То есть все просто?

Алгоритм из трех шагов:

- Выбираем метрику;
- Даем модели на вход какие-то данные;
- Считаем метрику по результату использования модели.

И задача решена?

03

Компоненты системы применения модели

Система применения модели

04

Зоны воздействия для оценки эффективности

Входящие данные 1/6

Подход 1. Сабсэмплинг;

Подход 2. Перемешивание;

Подход 3. Дополнение;

Подход 4. Генерация;

Подход 5. Уничтожение.

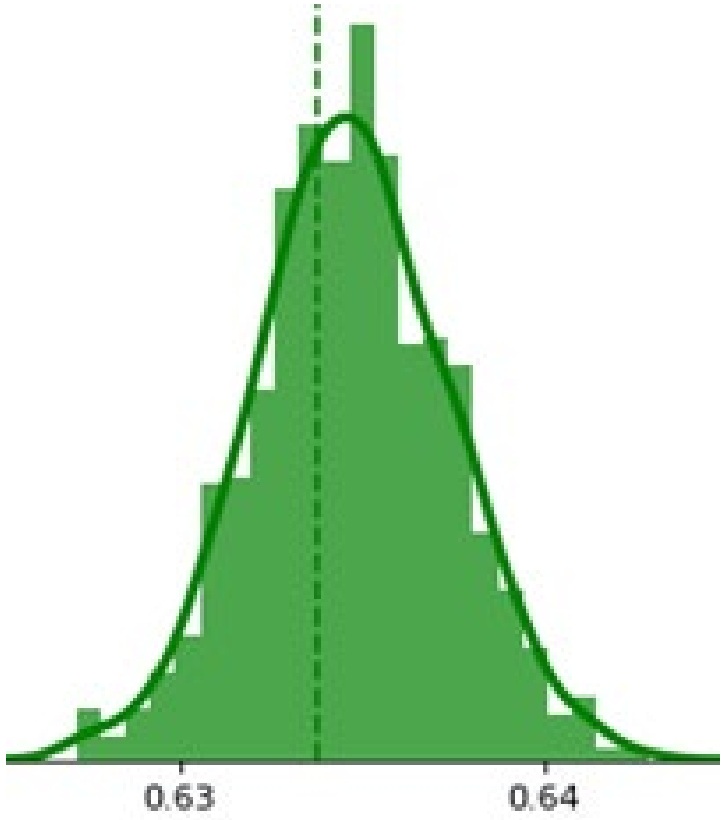
Входящие данные 2/6 . Сабсэмплинг

Повторяем расчет метрики многократно на подвыборках (например, на 80% исходной).

Точечный Gini равен 72,3%.

Но ни на одной из 10 000 подвыборок Gini не превысил 64,5%.

Оценка эффективности модели здесь основывается на математическом ожидании и доверительных интервалах метрики.



Входящие данные 3/6 . Перемешивание

Перемешиваем значения факторов в выборке.

Подаем эту выборку на вход модели.

Считаем метрику.

Если она улучшилась или почти не изменилась, то мы должны задать себе вопрос – а работает ли модель сама по себе?

Возможно используется внешний источник факторов.

Входящие данные 4/6 . Дополнение

Обогащаем выборку чем-то похожим (или не очень).

Подаем эту выборку на вход модели.

Считаем метрику.

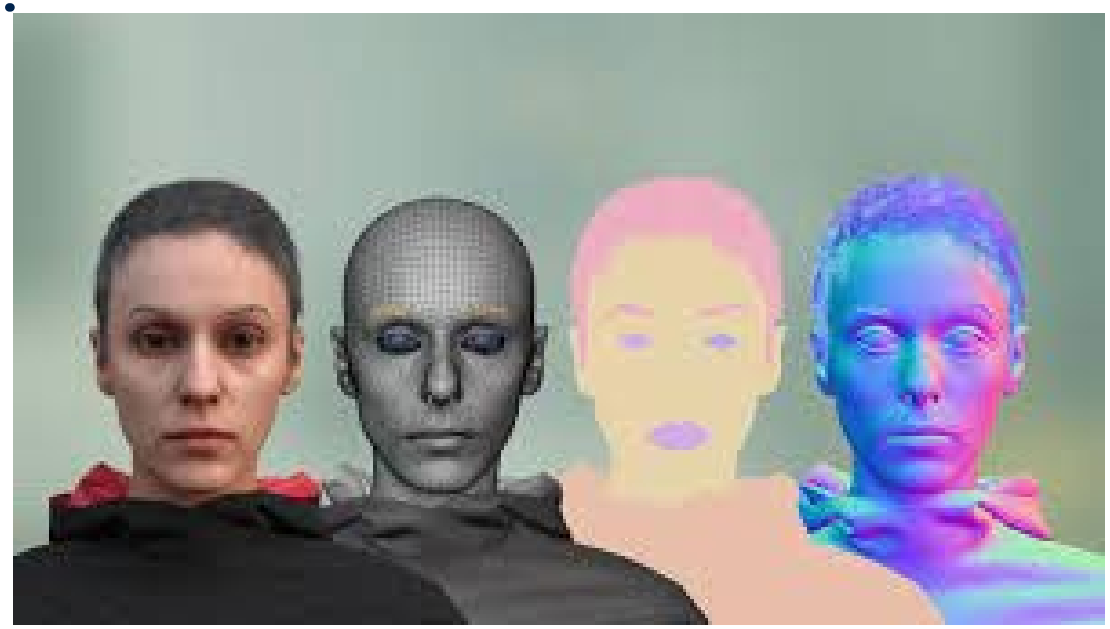
Если она улучшилась или почти не изменилась, это вызывает очень много подозрений.

Входящие данные 5/6 . Генерация

Обогащаем выборку или заменяем ее целиком синтетическими данными.

Подаем эту выборку на вход модели.

Считаем метрику.



Результат напрямую зависит от алгоритма генерации данных и остается на усмотрение аналитика.

Входящие данные 6/6 . Уничтожение

Уничтожаем часть значений переменных.

Подаем эту выборку на вход модели.

Считаем метрику.

Если она улучшилась или почти не изменилась, то мы должны задать себе вопрос – а работает ли модель сама по себе?

Возможно используется внешний источник факторов.

Модель 1/2

Подход 1. Меняем управляющие параметры;

Подход 2. Переобучаем модель.

Модель 2/2 . Меняем управляющие параметры

Единого подхода к изменению параметров нет.

Мы со своей стороны фиксируем все кроме одного,
а один выбранный меняем в широком диапазоне.

Интерпретация почти всегда
экспертная, но стандартный
проблемный кейс это:

мы ждем от параметра одного
поведения, а получаем другое.

Например, увеличивая уровень
одобрения, уровень риска снижается.

Или меняем random_seed и получаем
прирост качества модели на 5 п.п.

Результат модели

На практике практически не воздействуем на этот этап.

Немногие исключения:

1) Замена результатов модели на константу для проверки эффективности процесса без нее;
2) Изменение результатов модели для проверки системы контроля качества данных.

Применение модели

На практике- практически не воздействуем на этот этап.

Единственное исключение - переход от статистических метрик оценки эффективности к стоимостным.

Подход правильный, но требует активной проработки системы оценки доходов и расходов.

Комбинируем зоны влияния и подходы

Объединяя вышеупомянутые подходы, мы приходим к новым:

- Проверка эффективности модели на кроссвалидации (сабсэмплинг данных + переобучение модели).
- Стресс-тестирование модели

(изменение управляющих параметров+ переход к стоимостной оценке эффективности)

- Оценка уровня модельного риска

(сабсэмплинг/уничтожение данных + переход к стоимостной оценке эффективности)

И многие другие.

настоящие
возможности

росбанк

30 лет