Классификатор для брендирование такси

Котюшев Михаил Юрьевич

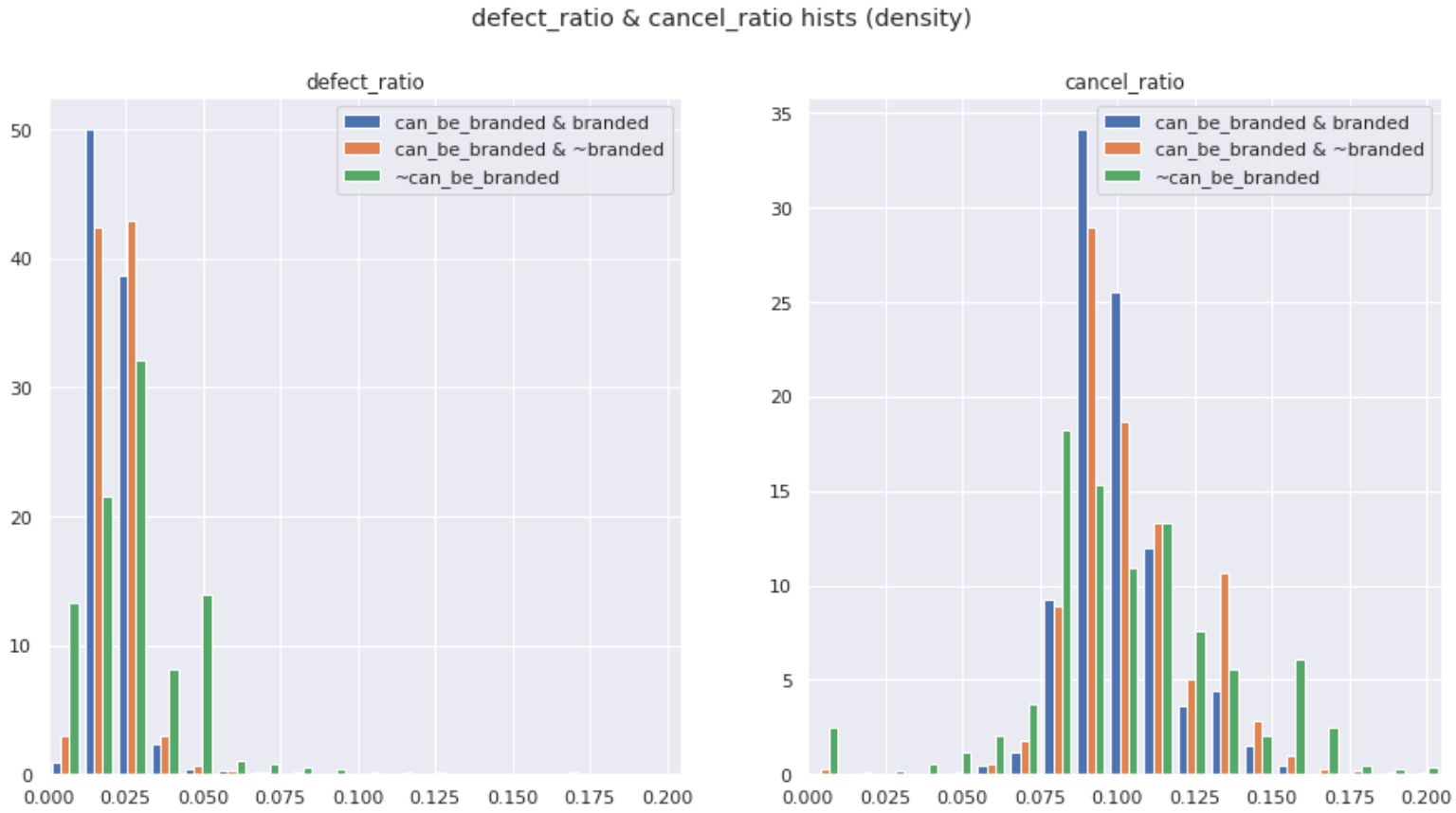
Практический Data Science, ШАД

Что будем оптимизировать?

Давайте посмотрим, какую долю занимают следующие подмножества:

- Машины, классифицируемые как брендируемые: ~ 79% от общего числа
- Брендированные машины: ~ 15% от общего числа и ~18% от брендируемых
- Не брендированные машины: ~ 85% от общего числа и ~82% от брендируемых

Взглянем теперь, как распределен defect rate внутри этих подмножеств:



Видно, что среди не брендируемых машин есть такие, Defect Rate которых ниже, чем у брендируемых. Их можно перетянуть в брендированные.

Давайте оптимизировать Defect Rate, не снижая (увеличивая) Share of Voice.

Как будем оценивать?

Допустим, мы придумали новый классификатор. Для подсчета метрик необходимо посчитать, сколько машин из положительно отклассифицированных будет действительно брендировано. Воспользуемся для этого следующим предположением:

Предположение I: доля машин, которые будут забрендированы из всех брендируемых останется неизменна при замене классификатора *для моделей, которые раньше были не брендируемы, а теперь стали*.
Исходя из этого предположения, оценим количество бренированных машин с новым классификатором так:

$$N(Б \ \& \ New = 1) =$$
$$N(Б \ \& \ New = 1 \ \& \ Old = 1) + N(New = 1 \ \& \ Old = 0) * (N(Б) / N(Всего))$$

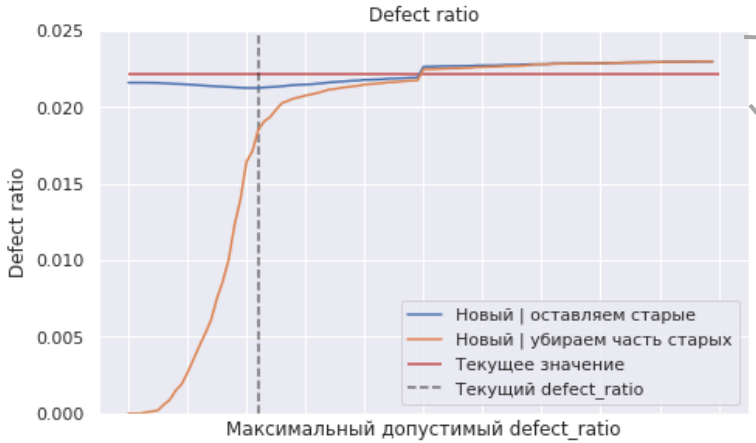
где N(Б & New = 1) -- прогноз на количество брендирований, N(Б & New = 1 & Old = 1) -- число брендирований среди пересечения классификаторов (знаем наверняка), N(New = 1 & Old = 0) -- общее число машин, добавленных в брендируемые новым классификатором, N(Б), N(Всего) -- общее число брендированных и общее число машин соответственно.

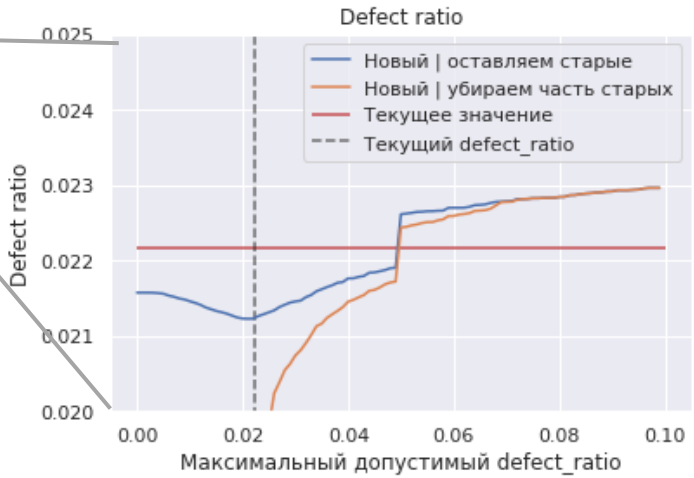
Также для подсчета метрики Share of Voice нужно знать общее число брендированных машин всех аггрегаторов в нашей локации. Заменим Share of Voice на Share of Branded в следующем предположении:

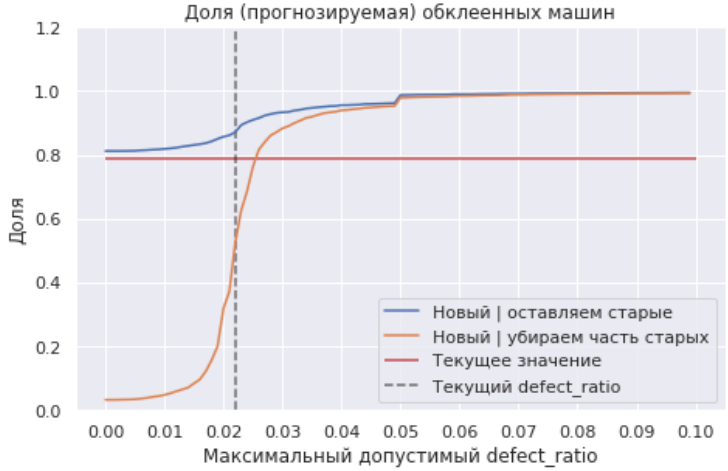
Предположение II: общее число брендированных машин всех аггрегаторов в нашей локации не поменяется за время внедрения. Таким образом, при росте Share of Branded будет расти и Share of Voice и можно оптимизировать Share of Branded, который можно посчитать из данных, классификатора и предположения 1.

Решение & внедрение

Поступим просто: зафиксируем порог Defect Rate, выбирая для брендирования модели с Defect rate меньше этого порога. Пробежимся по значениям порога в разумных пределах и выберем такой, который доставляет оптимум (в некотором смысле) нашим занчениям метрик.







Минимум среднего Defect Rate по брендированным машинам (синяя кривая на графике 1) здесь появляется по очевидным причинам -- прибавляя к выборке числа заведомо меньше среднего, уменьшаем это самое среднее.

Для старых партнеров: точка минимума среднего DR (max DR == ~0.02) будет являться решением в случае, когда мы выбираем по отношению к старым уже обклееным машинам стратегию "оставить". Это позволит нарастить метрику Share of Branded на ~0.05 и незначительно уменьшить средний Defect Rate на ~0.001 только за счет партнеров из таблицы (тех, кому старый классификатор запрещал брендироваться).

Для новых партнеров: кривые для стратегии "убираем" на самом деле моделирует поведение метрик для новых партнеров -- для них оптимальная точка лежит заведомо правее точки пересечения красной и желтой кривых на графике Share of Branded, так как мы не хотим уронить эту метрику ниже текущего значения Поэтому для новых партнеров будет использоваться значение max DR == ~0.025.

Обсуждение

Что сделано:

- Классификатор с двумя группами внедрения: для старых партнеров позволяет улучшить метрики Share of Brand с ~0.79 до ~0.85 и средний Defect Rate по брендированным машинам с ~0.0224 до ~0.0212, для новых партнеров: ~0.79 -> ~0.81 и ~0.0224 -> ~0.0200 соответственно.
- Классификатор получился лояльный к старым партнерам -- можно обойтись без сняния брендирования с уже обклееных машин, поэтому экономика партнеров не должна пострадать.

Что хорошо бы сделать ещё:

- Не использован файл partners.xlsx -- можно было бы проанализировать какие-то другие партнерские метрики
- Не расписан процесс внедрения по времени.