Соревнование: Digital Reputation

Николай Скачков, ММП 517

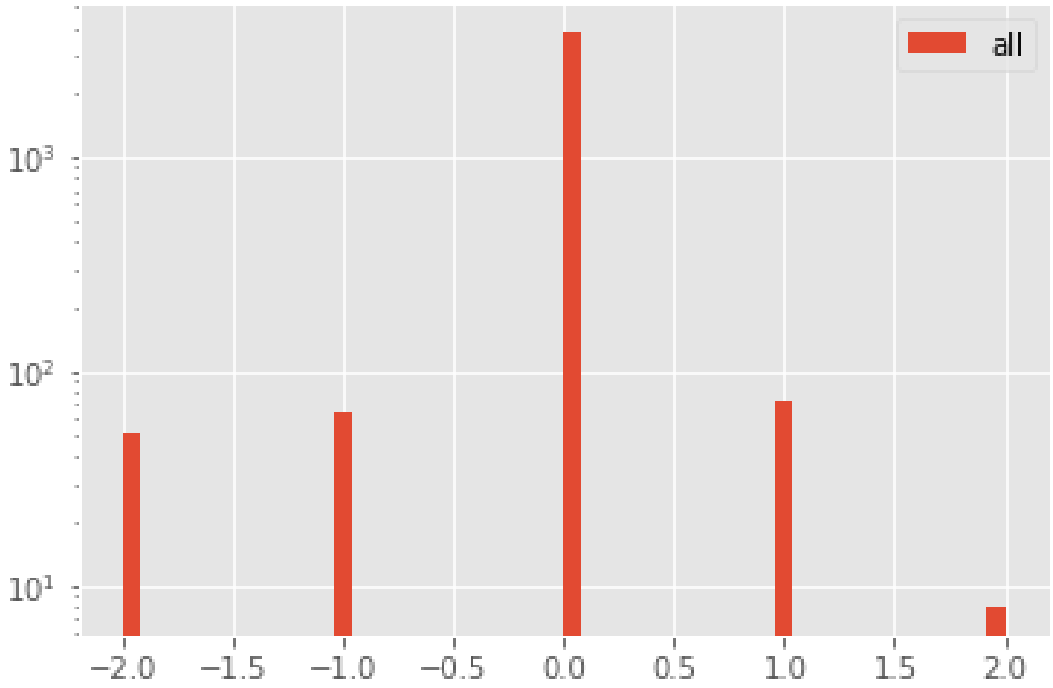
21 октября 2019 г.

Общий ход решения

Основные этапы решения:

1. Визуализация данных и преобразование признаков;
2. Отбор признаков;
3. Выбор модели по кроссвалидации/холдауту;
4. Ансамблирование;

Визуализация и преобразования признаков



Гистограмма 3-го признака в X1

1. Находим в X1 категориальные признаки и разбиваем их на бинарные;
2. К непрерывным признакам добавляем их логарифм.

Визуализация и преобразования признаков

Таблица X2:

1. Содержит огромное количество событий, привязанных к id;
2. Создаём разреженную матрицу id x unique_values.
3. Делаем PCA в 10 главных компонент.

Отбор признаков

- Учим бустинг на X1, X2 и X3. Получаем:

Данные обучения	Средний ROC AUC
X1	0.58
X2	0.50
X3	0.49

- Как видно, данные X2, X3 фактически бесполезные для обучения. Мы не будем добавлять их в обучение.

Выбор моделей

- учим модель логистистической регрессии для каждой из 5 задач;
- учим модель логистистической регрессии с предварительным отбором признаков с помощью L1 лог. регрессии для каждой из 5 задач;
- учим модель бустинга для каждой из 5 задач;
- Результаты:

Данные обучения	Av ROC AUC	Лучшая в
LogReg L2	0.58	-
LogReg L2 and L1	0.60	2, 3, 5
XGBoost	0.59	1, 4

Эксперименты с NN

- Хочется использовать информацию из теста в обучении.
- Учить автокодировщик плохо, так как он нечестно использует информацию из теста.
- будем учить Domain Adversarial модель, которой в своём внутреннем представлении запрещается выучивать что-то специфичное для трейновой выборки.
- внутренне представление учим размера 10 ли 100.
- DA заключается в дополнительному лоссе, который по скрытому представлению пытается предсказать из какого домена пришёл объект. Сеть учится обманывать этот лосс.

Эксперименты с NN

- Обучаем методом градиентного спуска. Оптимизатор Adam.
- Результат ROC AUC на holdout: 0.57
- Качество улучшить не удалось.
- Причина неудачи скорее всего в том, что для выучивания векторного представления в модель добавляется достаточно много параметров и данных не хватает для обучения качественной модели.

Итоги

- Отбор и преобразование признаков значительно улучшило результат.
- Сильно улучшило результат применение разщных моделей в разных задачах.
- Из X2, X3 не удалось вытянуть полезной информации.
- Использование теста в NN не дало прироста из-за малого размера обучающей выборки и переобучения.