Bank Marketing Research

Артём Уткин, Кирщин Иван

Ноябрь 2023

Содержание

1 Постановка задачи

Задача состояла в том, чтобы предсказать на основе данных, полученных португальским банком при массовом обзвоне клиентов, подпишется ли клиент на срочный депозит (переменная y).

Данные: Bank Marketing

2 Библиотеки

В процессе работы для загрузки датасета с UC Irvine Machine Learning Repository использовалась библотека ucimlrepo, для разведочного анализа данных использовались библиотеки pandas и numpy, для работы с моделями машинного обучения - scikit-learn.

3 EDA

Разбили датасет на тренировочную и тестовую выборку в соотношении 80 к 20. Последовательно обработали каждый имеющийся признак, проверили на незаполненные поля. Если таковые имелись, то в некоторых случаях обрабатывали как отдельное значение ('contact', 'poutcome'), в некоторых заполняли наиболее часто встречающимся значением ('education'), а в некоторых исключали из выборки ввиду незначительного количества пропусков ('job'). Затем обрабатывали выбросы: либо преобразовывали их ('balance'), либо выбрасывали строки с соответствующими значениями ('campaign'), если их было совсем уж мало. Применили нелинейные преобразования к признакам 'age' и 'campaign', чтобы приблизить зависимость между значениями признака и средними значениями целевой переменной к линейной. Часть категориальных признаков обрабатывали при помощи One-Hot Encoding ('default', 'housing', 'loan', 'job', 'marital', 'contact', 'poutcome', 'education'), а остальные категориальные признаки при помощи Mean Target Encoding ('day of week', 'month') ввиду большого количества различных значений. Для того, чтобы в дальнейшем корректно использовать l2-регуляризацию применили стандартизацию к числовым признакам.

4 Построение модели

Для классификации использовали логистическую регрессию с l2-регуляризацией. При помощи Grid Search на кросс-валидации перебрали 49 значений обратного коэффицента l2-регуляризации и подобрали оптимальный, коим оказался $C = 1$. Обучили на тренировочной выборке логистическую регрессию со значением коэффицента регуляризации равным 1.

5 Результаты

Для измерения качества классификации использовали метрику ROC-AUC. На тестовой выборке значение получилось равным примерно 0.77. Т.к. значение метрики получилось пусть и заметно большим чем 0.5 но также и заметно меньшим чем 1, это не позволяет назвать получившийся классификатор отличным, но и совсем плохим он тоже не является.

6 Заключение

Посмотрев на получившиеся коэффициенты логистической регрессии, мы смогли определить самые важные признаки, сильнее всего влияющие на решение клиента. Ими стали результат предыдущей маркетинговой компании, указан способ связи клиента или нет, является ли он студентом и в каком месяце ему предлагали оформить срочный вклад.