

# Bank Marketing Research

Артём Уткин, Кирщин Иван

Ноябрь 2023

## Содержание

<b>1</b>	<b>Постановка задачи</b>	<b>1</b>
<b>2</b>	<b>Библиотеки</b>	<b>1</b>
<b>3</b>	<b>EDA</b>	<b>1</b>
<b>4</b>	<b>Построение модели</b>	<b>1</b>
<b>5</b>	<b>Результаты</b>	<b>1</b>
<b>6</b>	<b>Заключение</b>	<b>1</b>

## 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  $y$ ).

## 2 Библиотеки

## 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

## 5 Результаты

## 6 Заключение