Прогнозирование оттока клиентов

Автор: Хабибуллин Салих

1. **Цели и задачи проекта.**

Отток клиентов - важная задача для любой телекоммуникационной компании, так как конкуренция на привлечение новых клиентов достаточно высока.

Решается задача Бинарной классификации - то есть максимально точно определить относится клиент к оотоку или к не оттоку.

Решением проблемы по удержанию клиентов может быть предложение более удобного тарифа, или скидка на текущий тарифный план.

Это можно сделать прямым обзвоном, или с помощью уведомления от приложения в телефоне, или с помощью чата в мессенджере.

Анализ данных

Отток : 7.44 %

Не отток : 92.56 %

В процентном соотношения для дальнейшего исследования присутствует сильный дисбаланс.

In [3]:
```
from IPython.display import Image
Image(filename = "head).jpeg")
```

Out[3]:

```
data.head()
```

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var222	Var223	Var224	Var225	Var226	Var227	Var228	Var229	Var
0	NaN	NaN	NaN	NaN	NaN	3052.0	NaN	NaN	NaN	NaN	...	vr93T2a	LM8l689qOp	NaN	NaN	fKCe	02N6s8f	xwM2aC7IdeMC0	NaN	N
1	NaN	NaN	NaN	NaN	NaN	1813.0	7.0	NaN	NaN	NaN	...	6hQ9lNX	LM8l689qOp	NaN	ELof	xb3V	RAYp	55YFVY9	mj86	N
2	NaN	NaN	NaN	NaN	NaN	1953.0	7.0	NaN	NaN	NaN	...	catzS2D	LM8l689qOp	NaN	NaN	FSa2	Zl9m	ib5G6X1eUxUn6	mj86	N
3	NaN	NaN	NaN	NaN	NaN	1533.0	7.0	NaN	NaN	NaN	...	e4lqvY0	LM8l689qOp	NaN	NaN	xb3V	RAYp	F2FyR07IdsN7I	NaN	N
4	NaN	NaN	NaN	NaN	NaN	686.0	7.0	NaN	NaN	NaN	...	MAz3HNj	LM8l689qOp	NaN	NaN	WqMG	RAYp	F2FyR07IdsN7I	NaN	N

5 rows × 231 columns

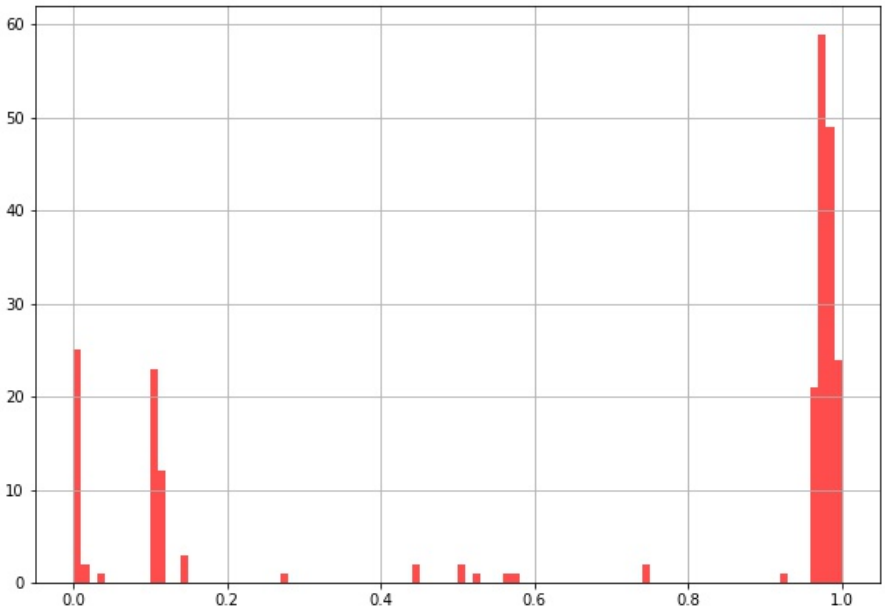
```
data.columns
```

```
Index(['Var1', 'Var2', 'Var3', 'Var4', 'Var5', 'Var6', 'Var7', 'Var8', 'Var9',
       'Var10',
       ...
       'Var222', 'Var223', 'Var224', 'Var225', 'Var226', 'Var227', 'Var228',
       'Var229', 'Var230', 'labels'],
      dtype='object', length=231)
```

В данных очень много пропущенных значений:

In [4]:
```
Image(filename = "features_and_NaNs.jpeg")
```

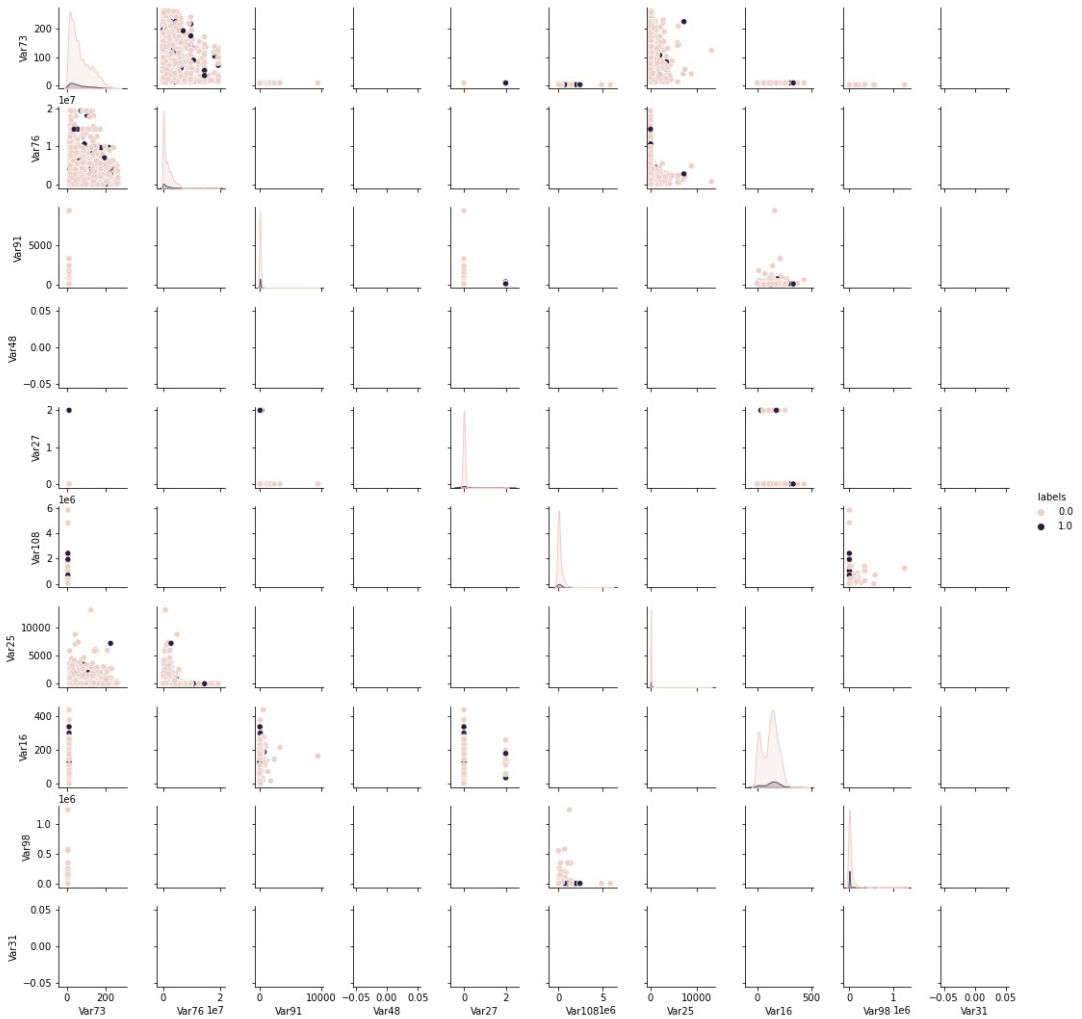
Out[4]:



In [5]:

```
Image(filename = "10_features.jpeg")
```

Out[5]:



2. **Методика измерения качества и критерий успеха.**

В данной задаче нужно как можно точно определять класс отток - за это отвечает метрика precision, но так как класс отток составляет всего ~7.5% от всего выборки, то есть присутствует явный дисбаланс классов.

Поэтому нужно взять метрику ROC-AUC, которая позволит учитывать несбалансированность классов, долю идентифицированного оттока и долю верных предсказаний с классом отток.

Процесс тестирования полученной модели нужно проводить на отложенной выборке (новых данных), также можно использовать AB-тестирование.

Критерий успеха - значительно увеличение прибыли относительно уровня, когда не производится работа по удержанию клиентов, или незначительное снижение прибыли относительно уровня, когда есть отток, но была произведена отличная работа по удержанию клиентов.

3. **Техническое описание решения.**

Были перепробованы разные модели решения:

- 1) Oversampling, Undersampling
- 2) Заполнение пропущенных значений у числовых признаков медианой
- 3) Для категориальных признаков LabelEncoder, OneHotEncoder.

- 4) RidgeClassifier, RandomForestClassifier, GradientBoostingClassifier
- 5) GridSearchCV
- 6) Отбор признаков по количеству пропусков

Pipeline:

- 1) Находим и не рассматриваем дальше полностью не заполненные признаки и одиночные признаки.
- 2) Полученные признаки разделяем на числовые и категориальные.
- 3) Отбираем признаки исходя из процентного обладания пропусками в этих признаках.
- 4) Заполняем средним числовые признаки, у категориальных - "missing".
- 5) Для итоговой модели используем библиотеку CatBoostClassifier.
- 6) Оцениваем качество по ROC-AUC на отложенной выборке, потом на тестовых данных.
- 7) Подбираемм лучшие параметры для CatBoostClassifier с помощью GridSearchSV и интуиции.

Экономическая модель.

(1 - per_churn) *N* tarif + per_churn *N* (percent_back * new_tarif[i] - zatrat)

- (1 - per_churn) - доля оставшихся клиентов
- N * tarif - базовая выручка, когда не было никакого оттока
- per_churn * N - доля клиентовв-оттока

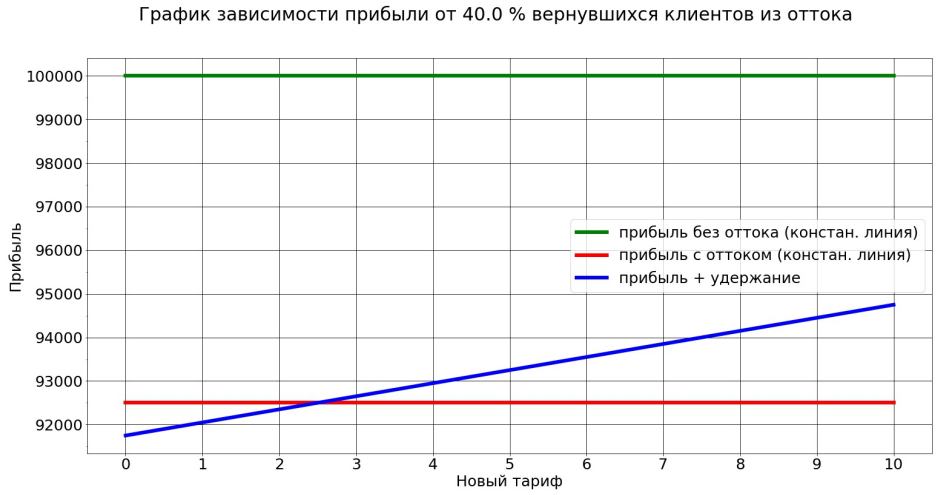
- per_churn *N* percent_back * new_tarif[i] - новая прибыль от клиентов оттока, которые передумали уходить
- per_churn *N* zatrat - затраты на людей из оттока, чтобы попытаться их удержать

Предположим, что прибыль от 1 человека в день и тариф - это одно и тоже.

- tarif - тариф
- new_tarif - новый тариф для тех людей, которые в группе оттока (чтобы их удержать)
- N - общее количесвто клиентов на старте
- per_churn - процент оттока

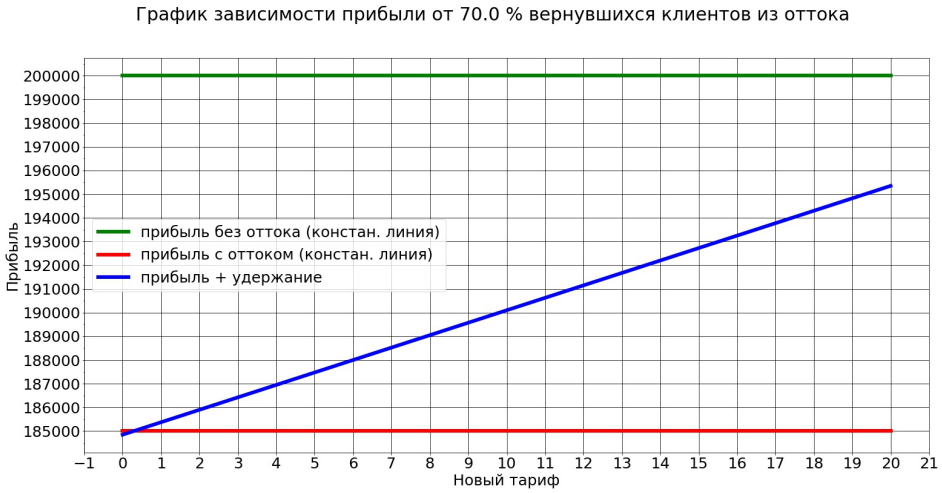
In [6]:
```
Image(filename = "picture_1.jpeg")
```

Out[6]:



In [7]:
```
Image(filename = "picture_2.jpeg")
```

Out[7]:



4. **Вывод о качестве модели.**

При увеличении качества модели на 1% или 3% будет несильное увеличение прибыли или вообще не будет.

Нужно уменьшить количество ошибок 1-ого и 2-ого рода, чтобы увеличить прибыль (экономический эффект).

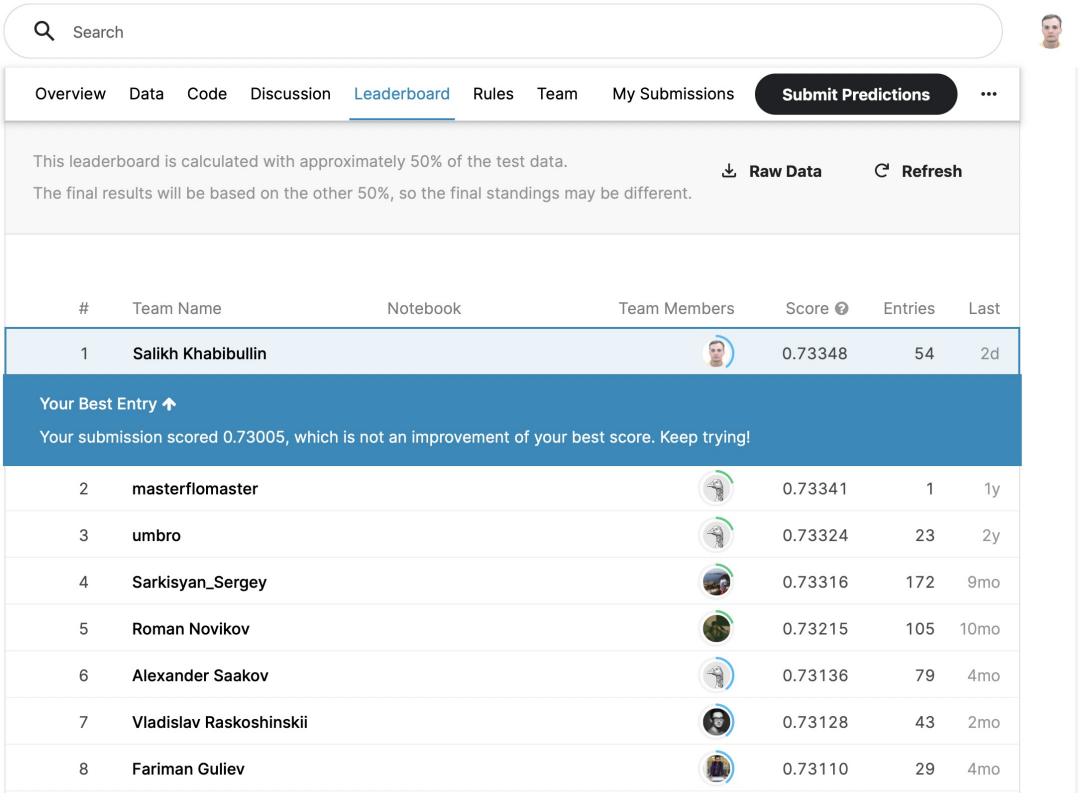
In [8]:
```
# Наиболее полезные признаки
Image(filename = "таблица).jpeg")
```

Out[8]:

	Importance	Feature
0	23.367443	Var126
1	5.191895	Var189
2	4.413075	Var199
3	3.423151	Var73
4	3.393323	Var113
5	3.204286	Var218
6	2.870235	Var192
7	2.754300	Var81
8	2.505554	Var74
9	2.258471	Var202

In [9]:
```
Image(filename = "Место).jpeg")
```

Out[9]:



5. **Итог работы.**

1) Нужна большая по размерам выборка.

2) Или данные, где будет намного меньше шума (шумовых признаков).

3) Иметь возможность тестировать на определённой группе людей в качестве эксперимента.

4) Чтобы более качественно получить прибыль от вложенных средств (получить большую отдачу),

нужно уменьшить количество ошибок 1-ого и 2-ого рода.

5) Оценивать прибыль с более реальными параметрами и дополнительными условиями и затратами,

что в итоге позволит судить о том, есть ли положительная динамика от вложения денег в улучшение качества модели.