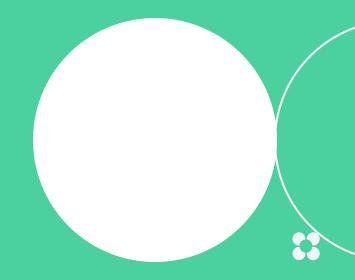
# Машинное обучение для предсказания дефолта по кредиту

Итоговая работа по курсу «Deep Learning»

Специализация «Data Scientist»



Малинина Елена Группа DSU - 1

# Постановка задачи



## Машинное обучение (Machine Learning, ML)

- форма искусственного интеллекта (ИИ), которая позволяет системе итеративно обучаться на данных, используя различные алгоритмы для описания данных и прогнозирования результатов. В данной работе рассматривается метод прогнозирование кредитоспособности клиентов банка Home Credit





## Актуальность задачи

Рост спроса на кредитование ведет к увеличению конкуренции на рынке и необходимость обработки большого количества данных. Такие условия обязывают почти полностью автоматизировать процессы с наименьшим участием человека.

Подобные процессы автоматизации также требуют гибкой настройки, адаптации, обучения на новых данных и построения более совершенных моделей.





### Описание проблемы

**Кредитный скоринг** представляет собой математическую или статистическую модель, с помощью которой на основе кредитной истории «прошлых» клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок.

Повышение доходности кредитных операций непосредственно связано с качеством оценки кредитного риска. В зависимости от классификации клиента по группам риска банк принимает решение, стоит ли выдавать кредит или нет, какой лимит кредитования и проценты следует устанавливать.

Основная задача – оценка риска



## Этапы решения задачи

- 1 Первичный анализ данных (EDA) и поиск аномалий
- 2 Оценка корреляции значений
- з Кодирование категориальных значений
- 4 Заполнение пропущенных значений
- 5 Конструирование признаков
- 6 Обучение логистической регрессии
- 7 Обучение CatBoostClassifier (модель градиентного бустинга)
- 8 Вывод



# Целевая метрика

Рассматриваемые нами метрики основаны на использовании следующих исходов:



- ТР истинно-положительное решение
- TN истинно-отрицательное решение
- FP ложно-положительное решение
- FN ложно-отрицательное решение

Тогда, точность и полнота определяются следующим образом:



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = rac{TP}{TP + FN}$$

F-мера представляет собой <u>гармоническое среднее</u> между точностью и полнотой.

$$F = 2rac{Precision imes Recall}{Precision + Recall}$$



# Анализ данных

Exploratory Data Analysis (EDA)



Был использован <u>датасет</u>, содержащий информацию о клиентах банка Home Credit Ноутбук доступен по <u>ссылке</u>

#### • Объем датасета:

307511 строк. Одна строка - представляет один кредит. В датасете указаны характеристики заемщика, такие как пол, возраст, место работы, семейный статус, уровень дохода, наличие недвижимости и машины и т.д, всего 122 Взяты также несколько дополнительных датасетов. Общий датасет составил 207 колонок.

#### Target:

является целевой бинарной переменной, где 0 - кредит погашен, 1 - кредит не погашен

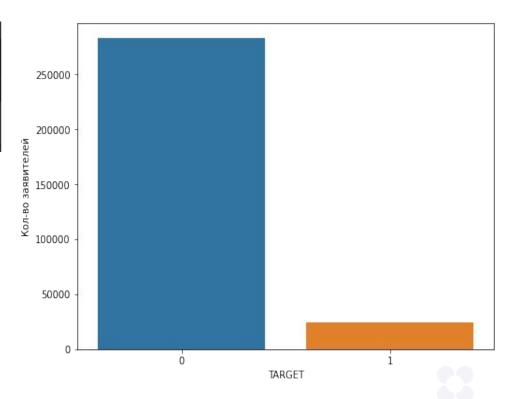


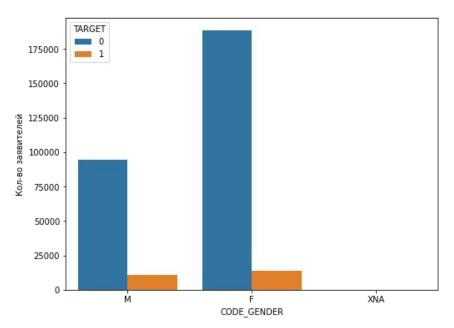


# Целевая переменная и распределение

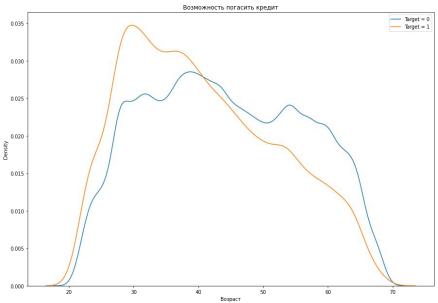
```
1 target_count = df_train['TARGET'].value_counts()
2 print('Knacc 0:', target_count[0])
3 print['Knacc 1:', target_count[1]])
Класс 0: 282686
Класс 1: 24825
```

Данные не сбалансированы. Это приведет к неправильному прогнозу в пользу 0-го класса

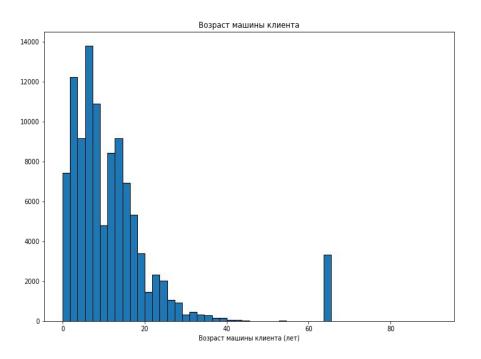


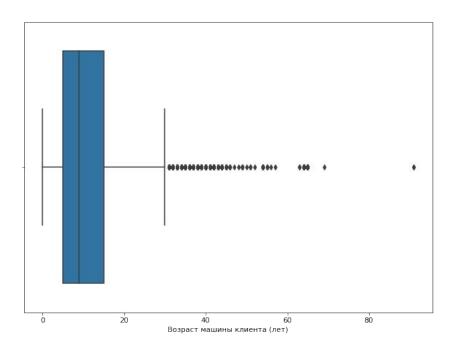


Рассматривается пол заемщика и чаще всего это женщины. Судя по графику проблемы с выплатой у женщин почти наравне с мужчинами, хотя мужчины берут кредиты реже. Также присутствует неизвестный класс XNA, значений в нем всего 4 шт. и их можно убрать.



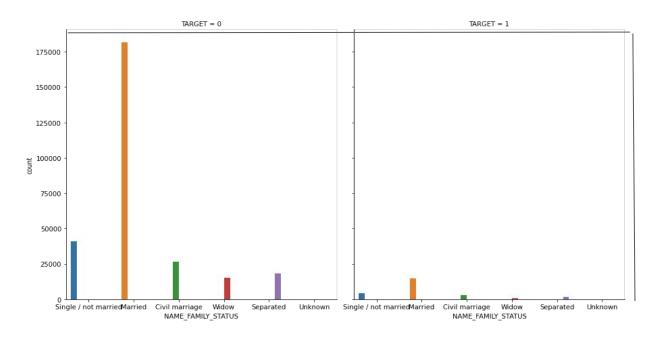
Разброс возраста от 20 до 70 лет, всплеск кредитования приходится в диапазоне от 25 - 45 лет. Можно увидеть, что проблемы с выплатой приходится примерно на тот же период жизни. При этом, чем старше человек, тем меньше вероятность не погасить кредит.





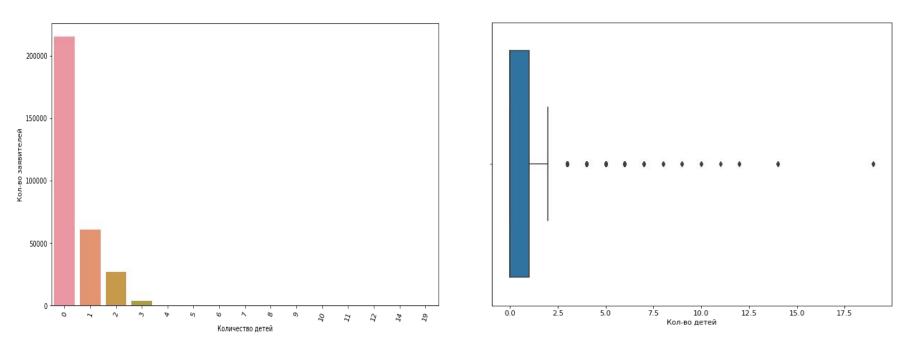
Возраст машины клиента имеет выбросы - их уберем





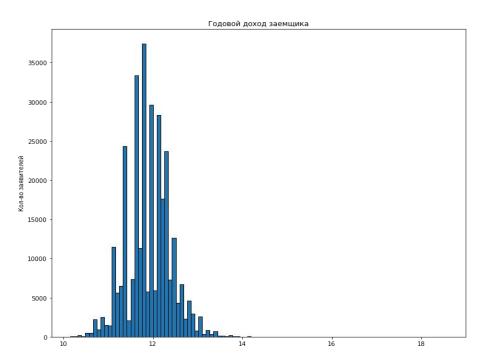
Семейный статус клиента может быть весомым при прогнозе, его тоже рассмотрим. Большинство женаты / замужем. Есть класс Unknown, там представлены 2 строки, его убираем.

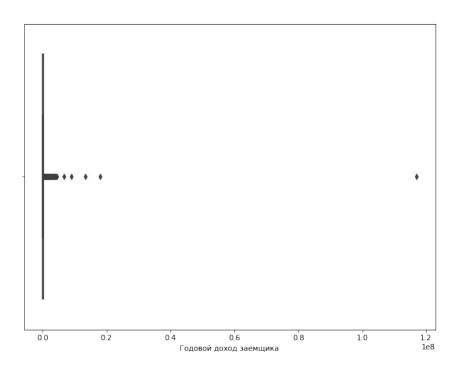




Большинство детей не имеют. На boxplot есть отметки, где больше 10 детей. Это вполне реальная ситуация ,но этих данных всего 8 строк, их тоже уберем, так как может быть перекос в данных.

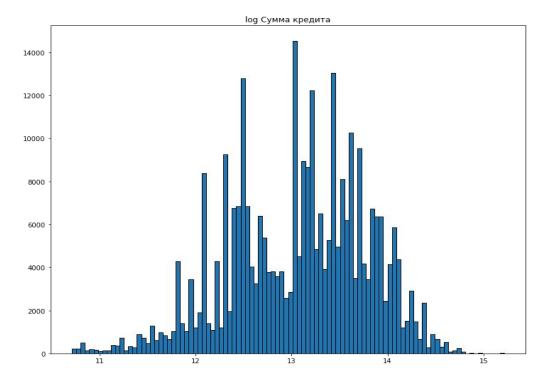






В boxplot годовом доходе заемщика есть значения превышающие 12 000 000. Их всего 3 шт. Эти данные лучше убрать из - за большого разрыва.

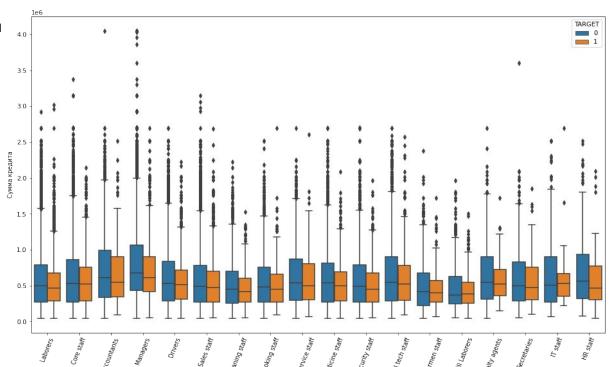


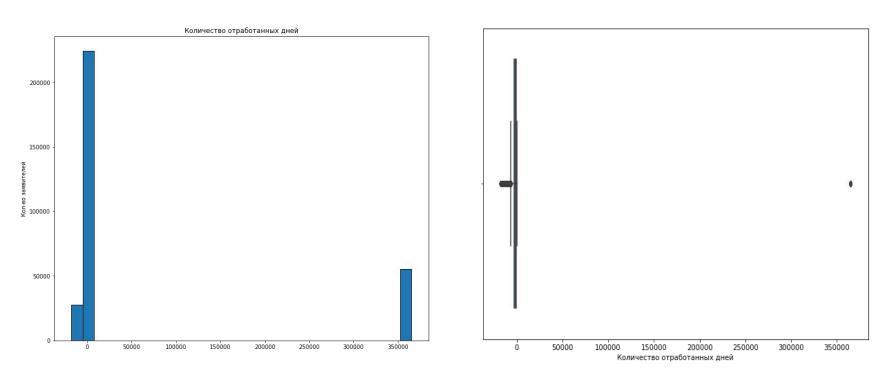


Кредиты на сумму > 400000 брали коммерческие партнеры в основном



Отношение типа занятости и суммы кредита, где люди чаще имеют задолженности.
Здесь присутствуют большие выбросы - их отсечем





Взглянем на количество отработанных дней. Есть значение в 365243 дня, оно равняется 1000 лет. Эти данные составляют 18% от всего дата фрейма, это достаточно много, поэтому эти значения заменяем на нули.



## Матрица корреляции

Среди предоставленных документов клиентом банку есть значительные зависимости, отберем самые значения с самой низкой корреляцией, а именно FLAG\_DOCUMENT\_6, FLAG\_DOCUMENT\_13, FLAG\_DOCUMENT\_16, FLAG\_DOCUMENT\_14,

FLAG\_DOCUMENT\_20

| TARGET -           | 1       | 0.0054            | 0.044             | -0.0027           | -0.0003           | -0.029            | -0.0015           | -0.0081           | -0.0043           | -0.0014            | -0.0042            | 0.00076            | -0.012             | -0.0096            | -0.0065            | -0.012             | -0.0034            | -0.008             | -0.0014            | 0.00022            | 0.0037             |
|--------------------|---------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| FLAG_DOCUMENT_2    | 0.0054  | 1                 | -0.01             | 5.9e-05           | 0.00081           | -0.002            | -8.9e-05          | -0.0019           | 0.00043           | -3.1e-05           | 0.00043            | 1.7e-05            | 0.00039            | 0.00035            | 0.00023            | 0.00069            | 0.00011            | 0.00059            | 0.00016            | 0.00015            | 0.00012            |
| FLAG_DOCUMENT_3    | 0.044   | -0.01             |                   | -0.014            | -0.19             | -0.49             | -0.021            | -0.47             | -0.098            | -0.0075            | -0.093             | -0.0012            | -0.02              | 0.0021             | 0.00059            | 0.033              | 0.0025             | 0.008              | 0.0094             | 0.0074             | -0.026             |
| FLAG_DOCUMENT_4    | -0.0027 | 5.9e-05           | -0.014            |                   | -0.0011           | -0.0028           | 0.00012           | -0.0027           | 0.00056           | 4.3e-05            | 0.00056            | -2.3e-05           | 0.00054            | 0.00049            | 0.00031            | -0.0009            | 0.00015            | 0.00082            | 0.00022            | -0.0002            | 0.0001             |
| FLAG_DOCUMENT_5    | -0.0003 | 0.00083           | -0.19             | -0.0011           |                   | -0.038            | -0.0017           | -0.037            | -0.0077           | 0.00059            | -0.0078            | 0.00032            | -0.0074            | -0.0067            | -0.0043            | -0.012             | -0.002             | -0.011             | -0.003             | -0.0028            | 0.0006             |
| FLAG_DOCUMENT_6    | -0.029  | -0.002            | -0.49             | -0.0028           | -0.038            | 1                 | -0.0043           | -0.092            | -0.019            | -0.0015            | -0.019             | 0.00079            | -0.018             | -0.014             | -0.0098            | -0.027             | -0.0044            | -0.024             | -0.0048            | -0.006             | -0.0057            |
| FLAG_DOCUMENT_7    | -0.0015 | 8.9e-05           | -0.021            | 0.00012           | -0.0017           | -0.0043           | 1                 | -0.0041           | 0.00086           | 6.6e-05            | 0.00086            | -3.5e-05           | 0.00082            | 0.00074            | 0.00048            | -0.0014            | 0.00022            | -0.0012            | 0.00034            | 0.00031            | 0.0002             |
| FLAG_DOCUMENT_8    | -0.0081 | -0.0019           | -0.47             | -0.0027           | -0.037            | -0.092            | -0.0041           | 1                 | -0.019            | -0.0014            | -0.017             | 0.0039             | 0.077              | 0.03               | 0.024              | 0.013              | 0.0053             | 0.0078             | 5.3e-05            | 0.0039             | -0.001             |
| FLAG_DOCUMENT_9    | -0.0043 | 0.00043           | -0.098            | 0.00056           | -0.0077           | -0.019            | 0.00086           | -0.019            | 1                 | -0.0003            | -0.0039            | 0.00016            | 0.00069            | 0.0014             | 0.0023             | 0.0085             | 0.0086             | 0.0019             | -0.0015            | -0.0014            | -0.001             |
| FLAG_DOCUMENT_10   | -0.0014 | 3.1e-05           | -0.0075           | 4.3e-05           | 0.00059           | -0.0015           | 6.6e-05           | -0.0014           | -0.0003           | 1                  | -0.0003            | 1.2e-05            | 0.00028            | 0.00026            | 0.00017            | 0.00048            | -7.8e-05           | 0.00043            | 0.00012            | 0.00011            | 8.7e-0             |
| FLAG_DOCUMENT_11   | -0.0042 | 0.00043           | -0.093            | 0.00056           | -0.0078           | -0.019            | 0.00086           | -0.017            | -0.0039           | -0.0003            |                    | 0.00016            | -0.0037            | -0.0034            | -0.0022            | -0.0063            | -0.001             | -0.0057            | -0.0015            | -0.0014            | -0.001             |
| FLAG_DOCUMENT_12   | 0.00076 | 1.7e-05           | -0.0012           | 2.3e-05           | 0.00032           | 0.00079           | -3.5e-05          | 0.0039            | 0.00016           | 1.2e-05            | 0.00016            | 1                  | 0.00015            | 0.00014            | -8.9e-05           | 0.00026            | 4.2e-05            | 0.00023            | -6.2e-05           | 5.7e-05            | 4.7e-0             |
| FLAG_DOCUMENT_13   | -0.012  | 0.00039           | -0.02             | 0.00054           | -0.0074           | -0.018            | 0.00082           | 0.077             | 0.00069           | 0.00028            | -0.0037            | 0.00015            | 1                  | -0.0032            | -0.0021            | -0.006             | 0.00097            | -0.0054            | -0.0015            | 0.043              | 0.0049             |
| FLAG_DOCUMENT_14   | -0.0096 | 0.00035           | 0.0021            | 0.00049           | -0.0067           | -0.014            | 0.00074           | 0.03              | 0.0014            | 0.00026            | -0.0034            | 0.00014            | -0.0032            | 1                  | -0.0019            | -0.0054            | 0.00088            | -0.0049            | -0.0013            | 0.031              | 0.002              |
| FLAG_DOCUMENT_15   | -0.0065 | 0.00023           | 0.00059           | 0.00033           | -0.0043           | -0.0098           | 0.00048           | 0.024             | 0.0023            | 0.00017            | -0.0022            | -8.9e-05           | -0.0021            | -0.0019            | 1                  | -0.0035            | 0.00057            | -0.0032            | 0.00085            | 0.028              | 0.0006             |
| FLAG_DOCUMENT_16   | -0.012  | 0.00065           | 0.033             | -0.0009           | -0.012            | -0.027            | -0.0014           | 0.013             | 0.0085            | 0.00048            | -0.0063            | 0.00026            | -0.006             | -0.0054            | -0.0035            | 1                  | -0.0016            | -0.0091            | -0.0024            | 0.082              | 0.0035             |
| FLAG_DOCUMENT_17   | -0.0034 | 0.00011           | 0.0025            | 0.00015           | -0.002            | -0.0044           | 0.00022           | 0.0053            | 0.0086            | -7.8e-05           | -0.001             | 4.2e-05            | 0.00097            | 0.00088            | 0.00057            | -0.0016            | 1                  | -0.0015            | -0.0004            | 0.026              | -0.000             |
| FLAG_DOCUMENT_18   | -0.008  | 0.00059           | 0.008             | 0.00082           | -0.011            | -0.024            | -0.0012           | 0.0078            | 0.0019            | 0.00043            | -0.0057            | 0.00023            | -0.0054            | -0.0049            | -0.0032            | -0.0091            | -0.0015            | 1                  | -0.0022            | 0.082              | 0.0063             |
| FLAG_DOCUMENT_19   | -0.0014 | 0.00016           | 0.0094            | 0.00022           | -0.003            | -0.0048           | 0.00034           | 5.3e-05           | -0.0015           | 0.00012            | -0.0015            | 6.2e-05            | -0.0015            | -0.0013            | 0.00085            | -0.0024            | -0.0004            | -0.0022            | 1                  | 0.035              | 0.0004             |
| FLAG_DOCUMENT_20 - | 0.00022 | 0.00015           | 0.0074            | -0.0002           | -0.0028           | -0.006            | 0.00031           | 0.0039            | -0.0014           | 0.00011            | -0.0014            | 5.7e-05            | 0.043              | 0.031              | 0.028              | 0.082              | 0.026              | 0.082              | 0.035              | 1                  | 0.007              |
| FLAG_DOCUMENT_21 - | 0.0037  | 0.00012           | -0.026            | 0.00017           | 0.00065           | -0.0057           | 0.00025           | -0.0015           | -0.0011           | 8.7e-05            | -0.0011            | 4.7e-05            | 0.0049             | 0.0023             | 0.00064            | 0.0035             | -0.0003            | 0.0063             | 0.00045            | 0.0075             | 1                  |
|                    | TARGET. | "LAG_DOCUMENT_2 - | FLAG_DOCUMENT_3 - | FLAG_DOCUMENT_4 - | FLAG_DOCUMENT_S - | FLAG_DOCUMENT_6 - | FLAG_DOCUMENT_7 - | -LAG_DOCUMENT_8 - | FLAG_DOCUMENT_9 - | FLAG_DOCUMENT_10 - | FLAG_DOCUMENT_11 - | FLAG_DOCUMENT_12 - | FLAG_DOCUMENT_13 - | FLAG_DOCUMENT_14 - | FLAG_DOCUMENT_15 - | FLAG_DOCUMENT_16 - | FLAG_DOCUMENT_17 - | FLAG_DOCUMENT_18 - | FLAG_DOCUMENT_19 - | FLAG_DOCUMENT_20 - | FLAG_DOCUMENT_21 - |
|                    |         | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               | FLAG               |



-0.2

## Кодирование категориальных переменных

Всего в датасете 16 категориальных переменных. Будем использовать стандартную библиотеку из Scikit Learn - LabelEncoder



- NAME\_CONTRACT\_TYPE
- CODE GENDER
- FLAG\_OWN\_CAR
- FLAG\_OWN\_REALTY
- NAME\_TYPE\_SUITE
- NAME\_INCOME\_TYPE
- NAME\_EDUCATION\_TYP
- NAME FAMILY STATUS
- NAME HOUSING TYPE
- OCCUPATION\_TYPE
- WEEKDAY\_APPR\_PROCESS\_START
- ORGANIZATION\_TYPE
- FONDKAPREMONT\_MODE
- HOUSETYPE\_MODE
- WALLSMATERIAL\_MODE
- EMERGENCYSTATE\_MODE



## Обработка пропущенных значений

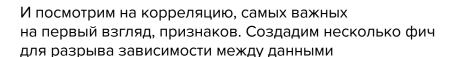
Пропущенных данных много. Справа показаны первые 20 колонок, где значений Nan больше всего. Воспользуемся Simple Imputer из библиотеки Scikit Learn, заполним переменные модой.



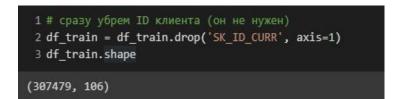


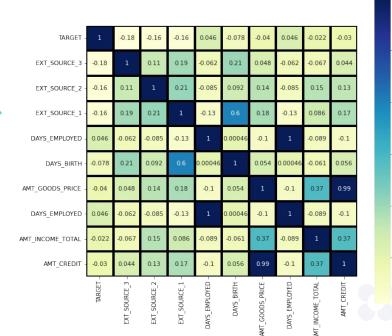
# Конструирование признаков (Feature Engineering)

Уберем ID клиента



- RATIO\_CREDIT\_PRICE\_% отношение кредита и стоимости покупок
- RATIO\_CREDIT\_ANNUITY\_% представление о сроках кредита
- □ RATIO\_GOODS\_INCOME\_% отношение зарплаты и стоимости покупок

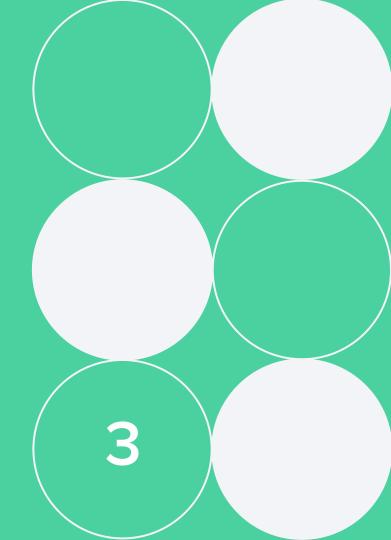




- 0.2

0.0

# Обучение алгоритмов

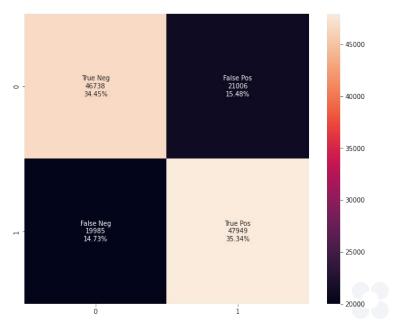


# LogisticRegression. Метрики

Для начала нормализуем данные с помощью <u>MinMaxScaler</u> и сбалансируем целевую переменную используя <u>библиотеку SMOTE</u>, где будем дублировать минорный класс. Делим данные (20% - тестовая часть). Так же посмотрим на матрицу ошибок

Качество модели на test: 0.6978802753578325 Качество модели на train: 0.6971264857966839

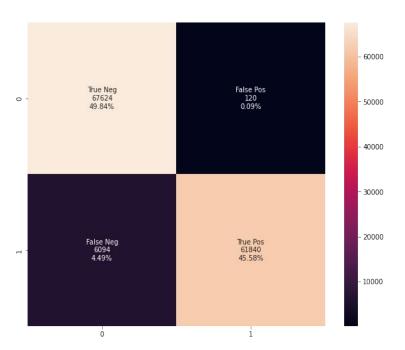
| Accuracy - 0.69<br>Recall - 0.7023<br>Precision - 0.6<br>ROC AUG - 0.696 | 943      |        |          |         |
|--|----------|--------|----------|---------|
| Р  | recision | recall | f1-score | support |
| 9  | 0.70     | 0.69   | 0.69     | 67744   |
| 1  | 0.69     | 0.70   | 0.70     | 67934   |
| accuracy   |          |        | 0.70     | 135678  |
| macro avg  | 0.70     | 0.70   | 0.70     | 135678  |
| weighted avg   | 0.70     | 0.70   | 0.70     | 135678  |
| weighten avg   | 0.70     | 0.70   | 0.76     | 133076  |



# CatBoostClassifier. Метрики

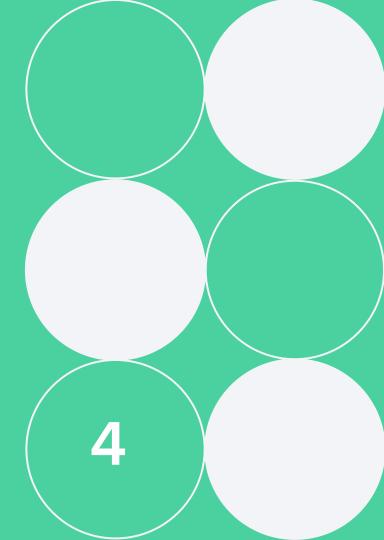
Качество модели на test: 0.9542003862085231 Качество модели на train: 0.9543623906677912

| 81      |              |   |                               |                                       |
|---------|--------------|---|-------------------------------|---------------------------------------|
|         |              |   |                               |                                       |
| ecision | recall       | f1-score  | support                       |                                       |
|         |              |   |                               |                                       |
| 0.92    | 1.00         | 0.96  | 67744                         |                                       |
| 1.00    | 0.91         | 0.95  | 67934                         |                                       |
|         |              |   |                               |                                       |
|         |              | 0.95  | 135678                        |                                       |
| 0.96    | 0.95         | 0.95  | 135678                        |                                       |
| 0.96    | 0.95         | 0.95  | 135678                        |                                       |
|         |              |   |                               |                                       |
|         |              |   |                               |                                       |
|         | 1.00<br>0.96 | ecision recall<br>0.92 1.00<br>1.00 0.91<br>0.96 0.95 | ecision recall f1-score  0.92 | ecision recall f1-score support  0.92 |





# Итоги обучения



# LogisticRegression. Результат

 Модель логистической регрессии не смогла показать хороший результат в решении этой задачи, точность в районе 70%.

| TRAIN                 | precis   | ion rec | all f1-s     | core support     |  |
|-----------------------|----------|---------|--------------|------------------|--|
| 0                     | 0.70     | 0.69    | 0.70         | 226366           |  |
| 1                     | 0.69     | 0.70    | 0.70         | 225890           |  |
| accuracy              |          |         | 0.70         | 452256           |  |
| macro avg             | 0.70     | 0.70    | 0.70         | 452256           |  |
| weighted avg          | 0.70     | 0.70    | 0.70         | 452256           |  |
| TEST                  | precisio | on reca | ll f1-sc     | ore support      |  |
| Ø                     | 0.70     | 0.70    | 0.70         | 56294            |  |
| 1                     | 0.70     | 0.70    | 0.70         | 56770            |  |
| accuracy<br>macro avg | 0.70     | 0.70    | 0.70<br>0.70 | 113064<br>113064 |  |
| weighted avg          | 0.70     | 0.70    | 0.70         | 113064           |  |
|                       |          |         | 2000         |                  |  |

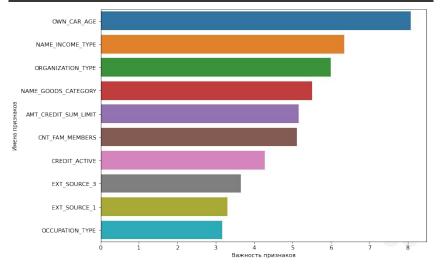


# CatBoostClassifier. Результат

 CatBoostClassifier показал хороший результат, переобучения нет

- На целевую переменную влияет:
  - OWN\_CAR\_AGE (кол-во лет машине владельца)
  - NAME\_INCOME\_TYPE (тип дохода клиента)
  - ORGANIZATION\_TYPE (организация, где работает клиент)

| TRAIN        | precis   | ion rec | all f1-s | core support |  |
|--------------|----------|---------|----------|--------------|--|
| 0            | 0.92     | 1.00    | 0.96     | 226299       |  |
| 1            | 1.00     | 0.91    | 0.95     | 225960       |  |
| accuracy     |          |         | 0.95     | 452259       |  |
| macro avg    | 0.96     | 0.95    | 0.95     | 452259       |  |
| weighted avg | 0.96     | 0.95    | 0.95     | 452259       |  |
| TEST         | precisio | on reca | ll f1-sc | ore support  |  |
| 0            | 0.92     | 1.00    | 0.96     | 56363        |  |
| 1            | 1.00     | 0.91    | 0.95     | 56702        |  |
| accuracy     |          |         | 0.96     | 113065       |  |
| macro avg    | 0.96     | 0.96    | 0.95     | 113065       |  |
| weighted avg | 0.96     | 0.96    | 0.95     | 113065       |  |
|              |          |         |          |              |  |
|              |          |         |          |              |  |



# Выводы



Исходя из этой работы можно сказать что, полученные результаты демонстрируют, что данная задача успешно решается методами классического машинного обучения

(2) Модель CatBoost лучше подходит для данной задачи, чем LogisticRegression

3 Из - за несбалансированной таргетной переменной нужно создавать синтетические данные

4

Гипотеза: возможно, добавление новых признаков (как анализ соц.сетей, например) повысит качество моделей

# Заключение



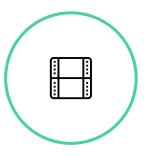
# 3 элемента



Произведен анализ исходных данных, выявлены и устранены значительные аномалии



Произведено сравнение результатов модели логистической регрессии и градиентного бустинга



Реализована модель ML, которая выявила определенные закономерности



