# Список файлов и что все они значат)

# 

# - Презентация - вообще актуальная преза у нас лежит в фигме, вот ссылка на нее https://www.figma.com/file/vWw2zxaLeIUkV9rjOeFxwT/%D0%9F%D1%80%D0%B5%D0%B7%D0%B5%D0%BD%D1%82%D0%B0%D1%86%D0%B8%D1%8F-MISIS-RYAZAN'?type=design&node-id=0%3A1&mode=design&t=0pq4GYZetst5UT1x-1

# - файл .ipynb - код с решением, выше представлена актуальная модель с препроцессингом, ниже нее идеи, которые мы также пробовали

# - corr.csv - список пар коррелирующих фичей, которые мы смогли найти, корреляция по модулю > 0.8

# - ann.csv - список id клиентов, которые мы обозначили как аномальные, используя автоэнкодеры

# Мысли, которые обязательно стоит учесть в питче:

1. Баланс классов, отработали
2. Аномальные пользователи - выявили
3. Нашли коррелирующие фичи - предоставим организаторам
4. Интерпретируемость модели
5. Поиск оттока даже на ранних стадиях
6. Блендинг = стабильность - банкам это и нужно

# Структура

Проблематика

Дисбаланс классов

Анализ фичей

Корреляция

Поиск аномалий

…

# Слайды:

## Проблематика

Что стоит знать для понимания проблемы:

1. Так как зарплатный клиент приносит банку больше прибыли, чем обычный клиент, за ним должно быть особое внимание, банк должен реагировать на его желание уйти и стимулировать остаться
2. Стимулировать остаться нужно вовремя, для этого стоит узнавать о возможном оттоке заранее
3. На плечи аналитиков ложится огромное количество фичей, с их большим количеством трудно работать, очень много неочевидных зависимостей

## Дисбаланс классов

Мы сразу же обратили внимание на дисбаланс классов. Мы пробовали использовать upsampling, class weights, второй способ оказался лучше. 1 класса в 27 раз меньше, поэтому мы пробовали дать его вес 27, а классу 0 - 1, однако лучше было распределение весов в соотношении 1:7. Также пришла абмициозная скажем так идея, разбить обучающую выборку на 27 подвыборок, в которой у нас одни и те же сэмплы с классом 1 и в том же количестве сэмплов с классом 0. Получилась довольно стабильная модель, однако она нам показалась слишком перепараметризирована, тем не менее, можно развить эту идею и сделать меньше моделей.

**Про RECALL**

## Анализ фичей

Большую часть хакатона мы провели за анализом данных. Мы заметили большое количество коррелирующих фичей. Из-за того, что признаки анонимизированы, мы не можем сказать, насколько очевидна эта корреляция, поэтому пары коррелирующих фичей мы предоставим организаторам, как результат нашей работы. Для обучения модели одну фичу из пары коррелирующих можно убрать. Может быть такое, что корреляция на первый взгляд не очевидна, это может дать аналитикам банка новые идеи для решения проблемы.

Также мы занялись поиском аномальных пользователей - они плохо влияют на стабильность модели, поэтому их также стоит убирать из обучающей выборки. Мы это реализовали при помощи автоэнкодера: его архитектура недостаточно параметризована, это сделано для того, чтобы модель не смогла выучить аномалии. Мы учим модель восстанавливать исходные данные, после обучения мы считаем MSE исходных данных и реконструкции, считаем аномальными клиентов, если их MSE лосс отличается больше чем 90 перцентиль.

Также мы поработали с отбором фичей. Для начала мы убрали около ста фичей, которые принимают только 1 уникальное значение. После этого мы использовали разные методы boruto и feature selection на основе катбуста. Все свелось к тому, что мы смогли отобрать 7% признаков практически без потери качества.

## Кластеризация

В какой-то момент получилось разбить сэмплы на кластеры на основе их эмбеддингов, полученных из автоэнкодера. На слайде показано разбиение на кластеры в пространстве пониженной размерности (понижали с помощью PCA). К сожалению, сильно не углублялись в эту идею, кластеры не дали видимого результата для модели, однако стоит больше проанализировать эту тему.

## Feature engineering

Также на 7% фичей, на которых мы обучались, мы смогли сделать новые фичи, путем применения арифметических операций, что дало видимый результат. Есть повод исследовать взаимосвязь этих фичей после деанонимизации

# Модели

Автоэнкодер - пробовали преобразовать наших пользователей в векторное пространство - получить эмбеддинги. Эмбеддинги подавали в катбуст - лучше не стало. Вероятно стоит глубже погрузиться в эту идею.

Также мы пробовали обучаться на MLP, пробовали использовать TabNet и AutoML - видимых результатов также не дало. Возможно ситуация может улучшиться после деанонимизации признаков.

Классические модели - random forest, knn, линейные модели, бустинги - все тестировалось. Лучшим оказался блендинг моделей CatBoost.

## Бизнес-идеи

На инференсе можно использовать информации о росте или падении вероятности оттока пользователя на каком-то промежутке времени с целью проверки гипотез о видах стимулирования клиентов.

Также, используя временные признаки, можно представить задачу как Survival Analysis, использовать соответствующие ей методы и модели для решения проблемы.

## Преимущества:

все на слайде

## 