# Цель:

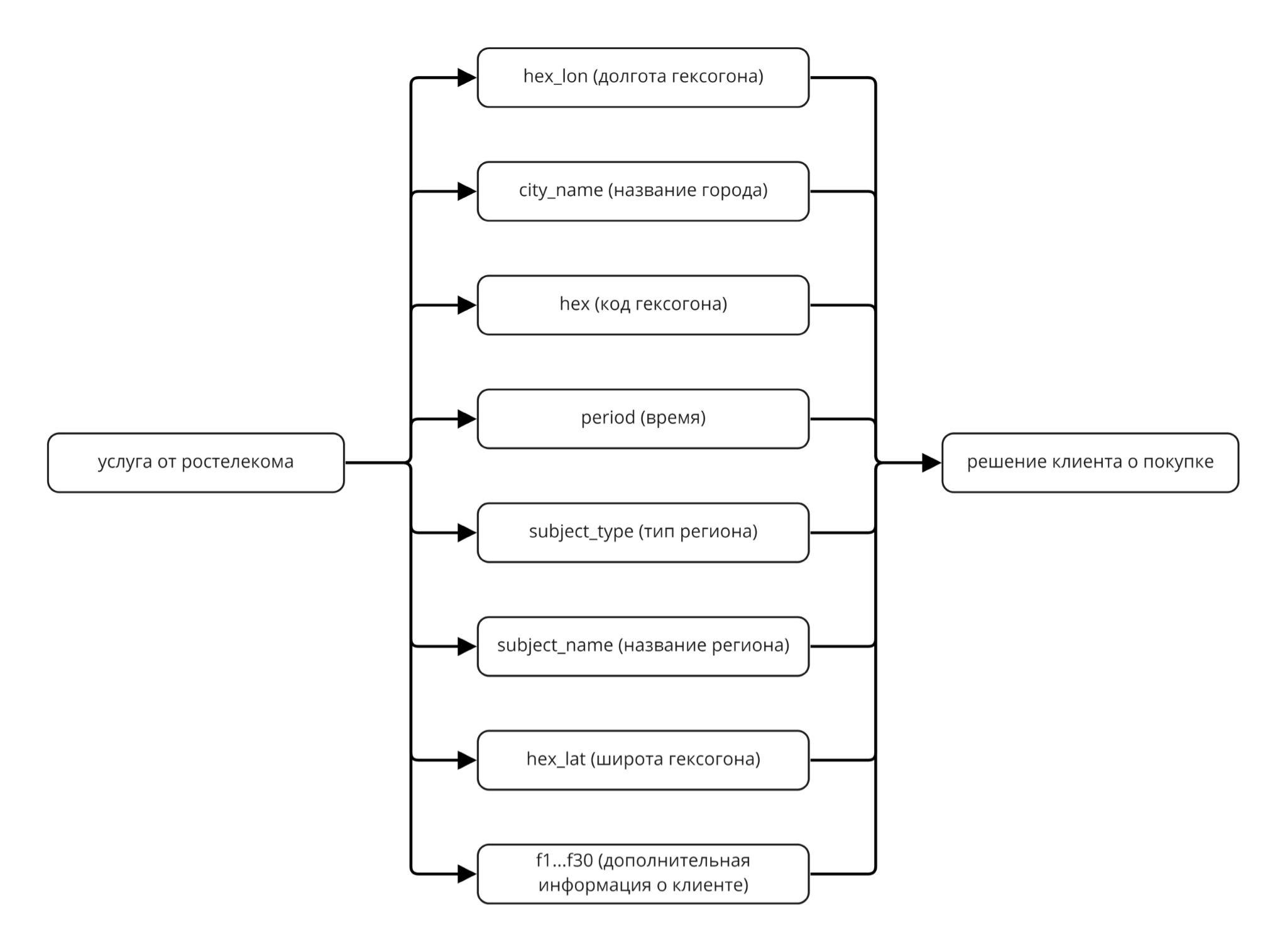
Разработка модели машинного обучения для предсказания склонности клиента к покупке услуги на основе предоставленных и открытых данных

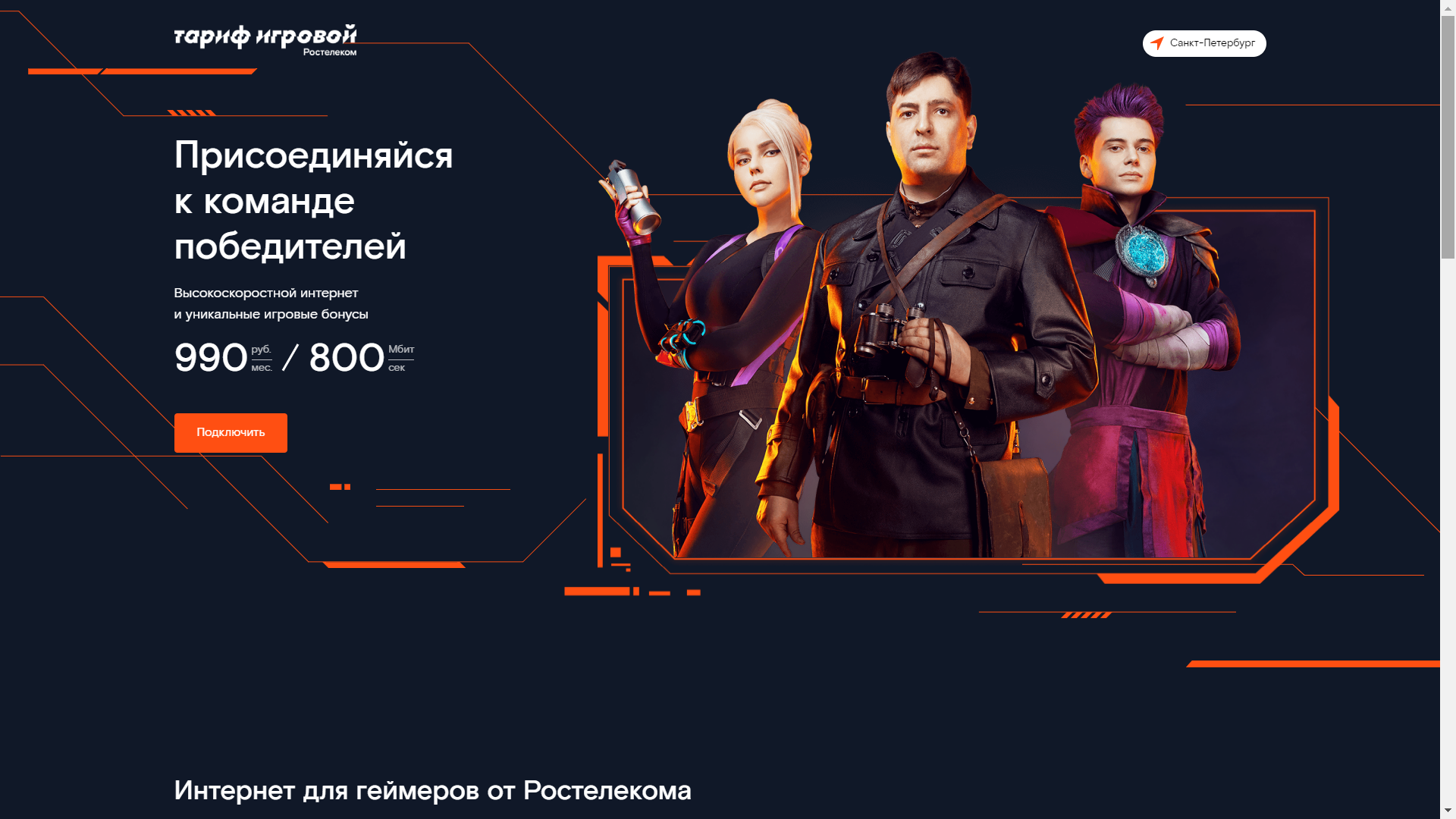
→→→ Нам нужно выявить информацию о клиенте, купит он или не купит. (1 или 0)

# Продукты Ростелеком:

* Домашний и мобильный интернет
* ТВ приставка
* Видеонаблюдение

# Визуализация задачи:





# Данные:

Регион, город, возраст, пол, история платежей, экономическая статистика, социальные и демографические показатели, курсы валют и фондовых рынков.

# Поля:

* Label – целевая метка
* period – год и месяц
* subject\_type – тип региона
* subject\_name – название региона
* city\_name – название города
* hex – код гексогона
* hex\_lat – широта гексогона
* hex\_lon – долгота гексогона
* f1…f30 – дополнительная информация о клиенте

# Идея:

При принятии решения на потенциального покупателя влияет множество факторов. Условно, для продажи домашнего интернета нужно учитывать физ или юр. лицо, возраст клиента, регион, время года, температуру окружающей среды, продолжительность светового дня.

Соответственно, нужно выявить закономерность(формулу) при стечении каких обстоятельств человек совершит покупку.

Тема нашей задачи — линейные модели

Теперь наша задача состоит в том, чтобы записать входные данные, записать выходные данные и подстроить сеть так, чтобы она сама подобрала параметры модели:

y=w1x1+…+wDxD+w0

Где y – целевая переменная (таргет), (x1,…,xD) – вектор, соответствующий объекту выборки (вектор признаков), а w1,…,wD,w0 – параметры модели.

# Задачи:

1. Разделить данные
2. Разработка модели машинного обучения для предсказания склонности клиента к покупке услуги на основе предоставленных и открытых данных

# Решение:

## Ход работы:

### 1 этап

Структурировать данные так, чтобы их понимала нейронная сеть и анализировать.

# Вопросы к первому чекпоинту?

* ~~Фичи 1 - 30 автогенерируемые? Они автоматически вставляются при подаче заявки или их прописывает пользователь?~~
* ~~Как понимать none? Она потеряна, не указана пользователем или специально не указана?~~
* ~~Значения полей city\_name в тестовых данных будут совпадать с значениями city\_name в тренировочных? Не будет ли новых городов?~~
* ~~Что такое hex в 6 столбике?~~

# Информация после первого чекпоинта:

* Рекомендательная модель улучшить спрос на конкретную услугу.
* Тариф “игровой”
* Построить рекомендательную модель
* Что за тариф, чем отличается, в чём специфика на рынке?
* пол, возраст,
* Внешние факторы обосновать???
* Вопросы с кейсодержателем?
* Сессия вопрос ответ в 10:00 27.08
* Потребители
* Кто из существующих допродажу примет
* Профили маркетинга.
* Чисто на игровой тариф, в тестовых данных тариф игровой.
* Про метрику завтра на сессии вопрос ответ
* Если странный город → метод НЕ ван хот энкодинг не даст ошибки
* работа обработка и подготовка данных и подтягивание. для обработки ван код енкодинг. тестировать пустые значения лучше вместо пустых
* One hot → урозерог отранжировать по численности даст прибавки. города представляют числа. Пропуски не учитывать.
* Кодировка с городами. Столбец А, Б. Можно 1 столбец 1 0. Как в пандас переменные. Каждый город закодировать. Много чего, чтобы как-то привести к нормальному виду.

# Выводы после первого чекпоинта:

* Информация о полкупателе: количество визитов/покупок, возраст клиента (время, прошедшее с первой покупки до текущего момента) и время, прошедшее с последней покупки.
* Важный критерий — есть ли оптоволоконный кабель в доме.

Процесс работы после чекпоинта

период 12 колонок с 1 и 11 0

субьект к основному дата фрейму и сбросить отделить дроп

захешировать субьект нэйм

город → ср. зарплата → сколько играет

код региона

сабджект тайп лучше бы говорил про тип региона — посёлок городского типа,

# 1 решение:

Убрать все текстовые данные.

Причина

В настоящее время в Российской Федерации существует более 150 тысяч населенных пунктов, более 18 тысяч городских и сельских поселений, более 2300 муниципальных районов и городских округов.

# 2 решение:

Заменить категориальные значения числовыми.

Идеи

Пройти циклом и очистить данные от улиц.

# Использованные материалы:

1. Работа с геоданными в Python и Jupyter // Proglib URL: https://proglib.io/p/rabota-s-geodannymi-v-python-i-jupyter-2021-03-22 (дата обращения: 26.08.2022).
2. Разбор маркетинга World of Tanks. Танки в маркетинге // ETDBOX URL: https://etdbox.ru/blog/6 (дата обращения: 27.08.2022).

# Сессия вопрос-ответ с кейсодержателем:

* Исторические данные до ноября включительно → подтянуть новости этого времени из других источников.
* Метрики → как считается метрика → когда обучим модель и будем формировать предсказание → нужно будет передать 5% людей, которые должны купить и отправить на оценку.
* Presigion на единичном классе → ищем людей которые купят → нулевой класс не купят.
* 1 - первая покупка? Или повторная? → разницы нет → клиент купил → абонентская плата. 1 - последние 3 месяца не было услуги, а потом появилась.
* На сколько в среднем покупается подписка? Которым интересна до конца.
* Поле период → год месяц, отсчётный период с первой покупки.
* Какой уровень у ростелекома сейчас? Чуть более сложное с точки зрения, несколько тысяч фичей.
* Расшифровка для фичей f1-30, специально с целью обезличивания, начисления клиента, трафик интернета, агрегированы за разные временные срезы, полгода, месяц.
* Можно ли использовать walid и test для кодирования.
* Метрики → топ 5 пресижина → тест датасет → отобрать 5% объектов, по мнению модели которой совершат покупку. Должны стоять единички. сабмит
* train test — не могу одновременно закодировать и тестовые модели? → Раздельно закодируем → разный набор фичей. Сгенерировать фичи.
* Что в презентации должно быть?
* За 3 часа до стоп кодинга
* Id - номера строкd

Есть два идентичных, но у одного не хватает поля. Мы дозаполним.

.

Критерии

Стабильности в стране

Сбои в текущего оператора

Строительство нового дома

Увеличение зр

Убрать лишний

Привести данные к 70/30

Цена рекламы

Просматриваемость YouTube / TV

Актуальность здорового образа жизни

# Второй чек-поинт

* ЦА, люди которые играют в игры от Ростелеком? Кто пользуются? Есть открытые источники, пол, возраст.
* Принятие решение. Создать модель, которая с большей вероятностью подтвердит. Нужно строить на внешних данных. Нужно понять, а какой ЦА идёт речь.
* Подойти с точки зрения маркетинга.
* Техническая часть →
* Если есть город и посёлок → в посёлок продавать не будем.
* Алгоритмы → лес и бустинги.
* Как будем обучать → все данные в траин → проверям тест.
* Алгоритм нейронные сети → готовые архитектуры → инстнет с полносвязной нейронной структурой → улучшить архитектуру → дропаут → регуляризацию, сколько дропаут 0,5 1 дропаут в конце.
* ПОИГРАТСЯ С АРХИТЕКТУРОЙ НЕЙРОННОЙ СЕТИ, ПРАВИЛЬНО ЛОС, ФУНКЦИЯ АКТИВАЦИИ РЕЛО И 2 ПОСЛЕДНИЙ СЛОЙ СИГМОИД. рело дропаут выход - 3 слоя. Надо увеличить. Мало параметров. КОЛИЧЕСТВО СЕТЕЙ. СТАТЬЯ ПРО ЦИФРЫ ОТ НУЛЯ ДО ЕДИНИЦЫ. ДИДЖИТС РЕКОГНИШН. АРХИТЕКТУРА СЛАБОВАТО. НАДО ПОКОВЫРЯТСЯ. БУСТИНГ → ЗАЧЕМ ЛЕС.
* НЕ СМОТРИТЕ НА ТОЧНОСТЬ, ДИСБАЛАНС В КЛАССАХ, НИКОМУ НЕ ПРОДАТЬ ТОВАР. ПРИВЕСТИ ТРЕНИРОВОЧНЫЕ ДАННЫЕ 70/30. 50/50 ИДЕАЛЬНО, В ДИСБАЛАНС — ВЫКИНУТЬ НУЛИ НЕ РЕМБЕЛ САЙЗ А ГОРОД. УМЕНЬШИТЬ ВЫБОРКУ И РАСПРЕДЕЛИТЬ. ЕДИНИЦЫ — ДУБЛИКАТЫ И ШУМ. →→→→ДОКРУЧИВАТЬ НОРМАЛИЗАЦИЮ И АЛГОРИТМЫ. F2 SCORE
* ЕСТЬ КУДА КОПАТЬ.
* ПРЕЗЕНТАЦИЯ → ИНФА В ТЕКСТОВОМ КАНАЛЕ. ПОДГОТОВИТЬ СКЕЛЕТ. КЛЮЧЕВЫЕ ОЦЕНКИ В КЕЙСЕ. СЛОВАМИ ПРОГОВОРИЛИ ЛУЧШЕ ВЫНЕСТИ НА СЛАЙДЫ. КАКИЕ СПОСОБЫ СТЕК ТЕХНОЛОГИЙ, КРИТЕРИЙ ОЦЕНКИ, ОБОСОНОВАНИЕ ВНЕШИНИХ ФАКТОРОВ + БАЛЛЫ.
* Следующий чекпоинт на 30 минут раньше.

Разбить бустинг на 2 слайда, через список параметры.

Указать график про каук будет полезен

На 2м результат. Два график как обучалась. То как росла точность. Цифрами какой точности достигли. Если есть цифры, как продвинулись в рамках хакатона. Алгоритм и техническая часть. И как достигли результата.

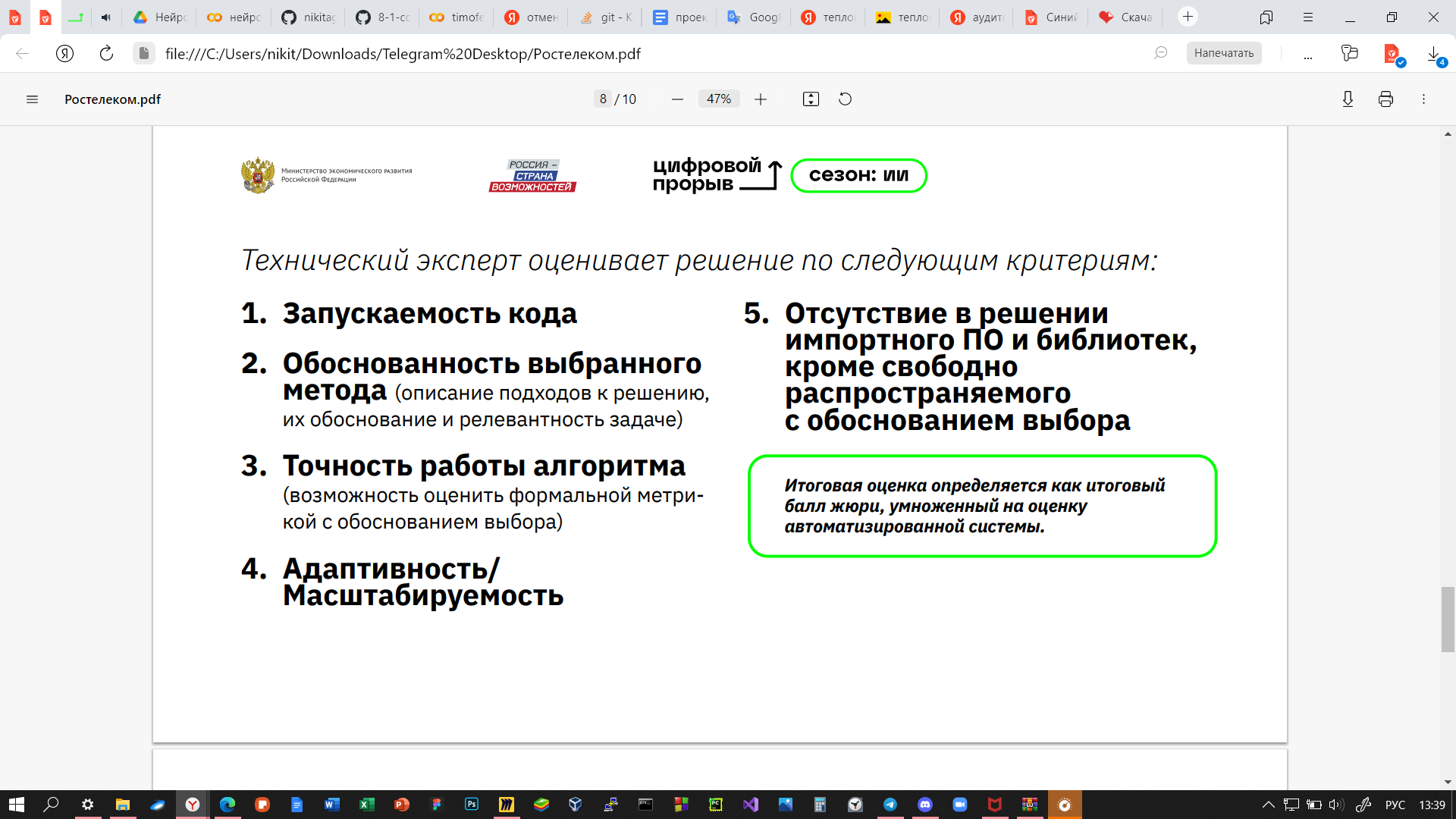
Точность не растёт в обучении или преобразования? По целевой метрике? Всё плохо.

В любом случае продемонстрировать

Чтобы бы мы потом сделали, но времени не хватило. Могли бы сделать это, это, чтобы повысить качество модели.

13 часов стоп кодинг. В 10 датасет. Комментарий, что куда выгружаем. На гитхаб. В 14 чек ап решений. В это время в дискорд. Посмотрим презентацию. Получим подтверждение.

Если вопросы, то в текстовом канале вопросы задавайте. Надо указатспикера с презентаций. В канале ТГ, как обозваться в зуме, чтобы на защиту допустили.



1. Программа была разработана в среде Google Colab. Код запускается полностью без ошибок и предупреждений.
2. Градиентный бустинг над решающими деревьями является одним из популярных методов решения проблемы классификации и выявлений аномалий, что подходит для использовании в представленной задаче.
3. Метрика ROC-AUC: 0.7470760534794361
4. Модель можно тренировать в дальнейшем с новыми данными и постепенно повышать ее точность.
5. Разработка велась в среде Google Colab на языке программирования Python с использованием open-source библиотек для анализа данных и машинного обучения.

Числовое представление категориальных данных может привести к повышению точности модели

Весомые данные

Скорость обработки

Применение многих внешних факторов

Параметры::

1. Catboost: 300 деревьев
2. LightGBM: 500 деревьев