1 слайд

Добрый день, уважаемая комиссия!

Сегодня мы рассмотрим **модель Одобрения займа на основе кредитной истории заемщика**

2 слайд

Для начала представлюсь

Меня зовут Адвокатов Дмитрий, я студент курса Data Science на платформе Skillbox.

На текущий момент я занят в немного другой сфере, но как раз по этому я и обратился за помощью на платформу. Курсы помогли мне начать и освоиться не только в работе с данными, но и в программировании как на Пайтон, так и на других языках.

Все это поможет мне в смене сферы деятельности и достижении моих личных целей. ….Если вам любопытно я хочу переехать в горы☺ … так же сменить направление своей профессиональной деятельности.

3 слайд

Но достаточно о банальном, давайте перейдем к тому, мы сегодня собрались.

Сейчас я постараюсь ***в меру?*** увлекательно и интересно презентовать вам свою работу по скорингу заемщиков:

Мы коротко поговорим о составе работы и чуть подробнее я задержусь на паре моментов, как мне кажется, заслуживающих внимания. Давайте начнем.

Целью работы было достижение достаточно точного предсказания при измерении его по метрике roc\_auc.

Для этого нам необходимо было решить несколько задач помельче:

Для начала нужно было понять как предполагается работать с данными, которые разбиты на несколько частей,

после этого предстояло выявить какие дополнительные признаки мы можем вытянуть из данных, которые будут более наглядны для машины и помогут ей в предсказании.

Как только у нас получилось определить наиболее оптимальную модель нам оставалось лишь создать удобный интерфейс для будущего потенциального пользователя.

4 слайд

Нельзя сказать, что в работе мы столкнулись с какими либо экстремальными испытаниями,

но были, скажем так, и ворота, пройти которые удалось не с ходу.

К примеру, необходимо было приспособиться к тому, что данные лежат в «разобранном» состоянии

Так же для нас это был первый проект с таким количеством данных – работа с ними потребовала как наращивания мощности локальной машины до оптимальных объемов памяти, так и оптимизации кода обработки данных.

5 слайд

со сбором в основном проблем не возникло – для работы на любезно предоставили замечательный инструмент как раз для решения этой задачи, мы его немного доработали и вуаля. Если конкретнее, то нам пришлось заменить библиотеку для чтения паркет-файлов, т.к. эта функция в «пандасе» работала нестабильно.

6 слайд

помимо этого мы сразу внедрили обработку в предоставленный инструмент для чтения и сбора данных.

**\*\*click!\*\***

Схема получилась достаточно несложная – на входе мы перечисляем функции-трансформерыв виде списка и передаем его в «читалку»,

**\*\*click!\*\***

внутри каждая функция оборачивается FunctionTransformer’ом и для удобства становится частью пайплайна. На выход же читалка отдает уже обработанный и в том числе агрегированный кусок датасета, который может состоять из от одной до 12 частей всего предоставленного датасета

7 слайд

На слайде вы можете наблюдать список функий, которые были написаны специально для обработки этого конкретного датасета, а так же в двух словах логику каждой из них.

Никакой некромантии применено тут не было – в основном мы старались максимально развить потенциал колонок с ключевой, как нам кажется, информацией, которая может повлиять на целевую переменную, т.е. одобрение нового кредитного продукта для заемщика.

К примеру, мы посчитали важным, сколько на текущий момент у заемщика неоплаченных займов, по скольким из них имеются просрочки.

Какой доход при этом заемщик приносит займодавцу

8 слайд

Еще мы посчитали важным отразить насколько часто пользователь закрывает займы раньше планируемой даты или насколько часто допускает просрочки по регулярным платежам.

В общем и целом как я уже отметил, мы старались развить потенциал «говорящих», как нам показалось колонок.

Помимо этих функций в пайплайн по обработке мы так же добавили некоторое количество технических функций – скажем так, административных – на слайде они выделены серым, т.к. не несут в себе ничего интересного, к примеру одна из таких функций просто «пришивает» целевую переменную к датасету, а другая удаляет ненужные колонки.

А вот о чем мы бы хотели рассказать немного подробнее, это о нашем агрегаторе:

9 слайд

**\*\*click!\*\***

перед нами стояла задача ужать данные на двадцати шести миллионах строк в сет на 3 миллиона и, желательно, сохранить при этом их информативность.

Как это сделать?

Просуммировать по пользователям? Не пойдет – колонки категориальные, мы исказили бы смысл

Оставить только самые последние записи? – тоже не пойдет, ведь смысл кредитной истории в, собственно, истории, а не в последней записи

Так мы пришли к нашему варианту – это своего рода вариация на тему зарекомендовавшего себя count encoder, только немного подругому.

**\*\*click!\*\***

Рассмотрим процесс на примере одной из колонок:

1. Возьмем колонку pre\_since\_confirmed для пользователя с id 0. У клиента в истории 10 записей,

**\*\*click!\*\***

7 из них имеют бинаризированную категорию 9, а так же категории 1, 0 и 12 – по одной записи на каждую.

**\*\*click!\*\***

1. Каждая из имеющихся категорий превращается в колонку, значение в которой будет отражать сколько раз эта категория встречалась в изначальной колонке для этого клиента. К примеру для клиента с id 0 после агрегации в датасете появится колонка pre\_since\_confirmed\_9\_agg и значение в ней будет – 7 – ровно 7 раз за имеющуюся у нас кредитную история этого клиента в изначальной колонке фигурировала категория 9.
2. По такому же принципу в колонках pre\_since\_confirmed\_ 0, 1 и 12 \_agg будут стоять единички.
3. Так минуточку – а откуда взялась колонка с нулем? Все просто – у нас есть и другие клиенты с другим набором категорий в изначальной колонке, по этому если у клиента с id 0 такая категория не появлялась, то уже у клиента с id 1 в агрегированной колонке будет красоваться единичка.

10 слайд

**\*\*click!\*\***

…а клиентов у нас 3 миллиона…

11 слайд

что же касается самой модели,

тут мы попробовали разные варианты, выявили ту,

что справляется лучше других, провели тюнинг по сетке,

зашили ее в интерфейс и двинулись дальше

12 слайд

в ходе исследований и тестов мы опробовали некоторое количество подходов к улучшению показателей проекта. К примеру

1. нам удалось сократить время работы модели за счет балансировки нулей и единичек – мы остановились на доле в 40% единичек в обучающем модель сете.
2. Соотношение на тестовой выборке, само собой осталось в неизмененном виде. Это дало нам ускорение обучения на 40% или в районе 145 секунд и практически без потерь в качестве.

13 слайд

в результате проделанной работы мы считаем, что достигли удовлетворительных результатов и смогли справиться с вызовами, представленными этим проектом, а именно:

1. При помощи наших кастомных и не очень обработок мы смогли подняться от бэйслановых 0,67% по метрике до искомых 0,75%+, т.е. перешли необходимый заказчику порог точности предсказаный. (*….хотя, думаю, бонуса мы, все же, не получим…)*
2. Так же спроектировали проект таким образом, чтобы все могло работать на вполне скромном железе
3. И, в дополнении, приняли меры, чтобы при необходимости обучить новую модель это можно было сделать достаточно быстро

14 слайд

На этом мой доклад заканчивается, спасибо за внимание! Теперь, если у вас остались или появились вопросы – пожалуйста задавайте их, а я буду стараться на них вам отвечать.