

Итоговый проект

Модель кредитного риска-менеджмента

Дорогие слушатели!

Вы завершаете курс Machine Learning Junior. Сейчас вы на финишной прямой. Осталось совсем немного — выполнить итоговый проект.

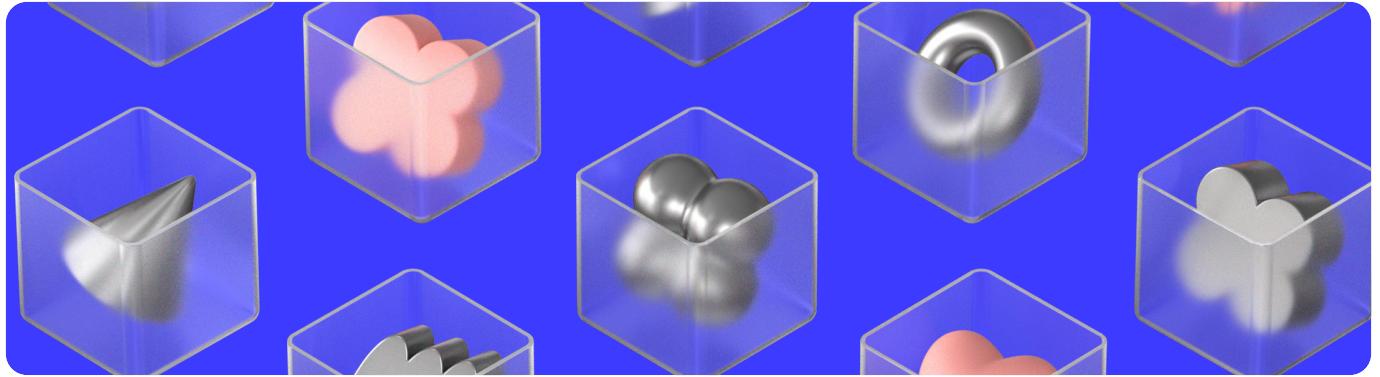
Вы уже освоили базовые алгоритмы машинного обучения, такие как регрессия, классификация и кластеризация, научились применять их на практике. Изучили базовые и важные алгоритмы машинного обучения. Научились проектировать простые нейронные сети на PyTorch, решать NLP-задачи, задачи предсказания для временных рядов и задачи обучения без учителя: кластеризовать данные, понижать размерности с помощью РСА и TSNE. Также вы умеете выстраивать пайплайны (от сбора данных до обучения модели на Python) и работать с PySpark.

Теперь вам предстоит довольно большая задача — обучить модель кредитного риск-менеджмента для банка. Выполняя её, вы проработаете все полученные на курсе знания и будете уверены в них на вашем первом рабочем месте в дальнейшем. Проект позволит вам закрепить следующие навыки:

- подбор и инжиниринг признаков для обучения модели;
- подбор лучшей модели для решения задачи классификации;
- подбор и настройка гиперпараметров модели;
- обработка больших массивов данных;
- интерпретация и визуализация результатов моделирования.

Проект имеет большую сложность и комплексность, поэтому будет отличным дополнением к резюме, подтверждающему ваши умения в области ML. Выполнение проекта — прекрасная возможность использовать все ваши знания и навыки, сделать важный шаг для карьерного роста и успешного будущего в области машинного обучения.

Уверены, у вас всё получится! Желаем успехов!



Как проходит работа над итоговым проектом

Перед вами подробное описание проекта. В нём есть всё, что вам нужно, чтобы справиться с поставленной задачей:

- Вводные данные по задаче.
- Техническое задание.
- Формат сдачи материалов.
- Информация о презентации итогового проекта.
- Критерии оценки.

Последовательно изучите каждую часть. Обращайтесь к актуальным для вас разделам в этом документе или к пройденным материалам курсов по мере выполнения задания.

Когда всё будет готово, отправьте проект через форму сдачи в курсе «Итоговый проект курса Machine Learning Junior».

Содержание

Вводные по задаче	5
Описание данных	6
Техническое задание	9
Формат сдачи материалов	11
Критерии оценки	11
Презентация итоговой работы	13

Вводные по задаче

Банки используют модели кредитного риск-менеджмента, чтобы понимать, насколько можно доверять клиенту в выполнении обязательств по договорам кредитования. Когда вы как клиент заполняете заявку на кредит или ипотеку, вас оценивают по модели кредитного риск-менеджмента. Банк может использовать разные сведения: например, о месте работы, возрасте, истории предыдущих погашений по другим кредитам в банках и кредитных организациях. На основе этой информации модель машинного обучения подсказывает кредитному менеджеру, стоит ли вам доверять запрашиваемую сумму денег.

С помощью такой автоматизации банк экономит время своих специалистов, чтобы они не искали и не агрегировали информацию по каждому клиенту для принятия решения о выдаче кредита. Это ускоряет время подтверждения заявки на кредит. Однако в отдельных случаях специалисты могут экспертно проверить решение модели, чтобы проаудировать её и выявить возможные слабые места.

Данный пример рассматривает только одну модель, которая учитывается в кредитном риск-менеджменте. Помимо неё используют и модели предсказания суммы кредита/займа, которую сможет оплатить клиент, и определения текущего рейтинга платёжеспособности клиента, у которого уже есть кредит. Модели нужны, чтобы спрогнозировать, какие клиенты могут выйти в просрочку, и предпринять какие-либо препятствующие выдаче кредита действия.

Проблема, которую предстоит решить

В рамках итогового проекта вы решите востребованную задачу — оцените риск неуплаты клиента по кредиту (дефолт).

Дефолт — неуплата процентов по кредиту или облигациям, непогашение займа в течение определённого времени t . Обычно дефолт считают свершившимся, если клиент не совершил выплату по кредиту в течение 90 дней.

Нужная модель позволяет банку или другой кредитной организации оценить текущий риск по любым выданным займам и кредитным продуктам и с большей долей вероятности предотвратить неисполнение кредитных обязательств клиентом. Таким образом, банк меньше рискует понести убытки.

Краткое описание задачи

Вам предстоит создать одну из моделей для оценки кредитного риска — предсказание выхода клиента в дефолт по кредиту.

Описание данных

Данные содержат информацию о различных атрибутах заёмщиков и кредитных продуктов: о клиентах, которые уже имеют кредиты, их кредитной истории и финансовых показателях. Каждая запись в датасете представляет один конкретный кредитный продукт, выданный конкретному заёмщику.

Атрибуты данных

- `id` — идентификатор заявки. Заявки пронумерованы так, что большему номеру соответствует более поздняя дата заявки.
- `rn` — порядковый номер кредитного продукта в кредитной истории. Большему номеру соответствует продукт с более поздней датой открытия.
- `pre_since_opened` — количество дней с даты открытия кредита до даты сбора данных (бинаризовано*).
- `pre_since_confirmed` — количество дней с даты подтверждения информации по кредиту до даты сбора данных (бинаризовано*).
- `pre_pterm` — плановое количество дней с даты открытия кредита до даты закрытия (бинаризовано*).
- `pre_fterm` — фактическое количество дней с даты открытия кредита до даты закрытия (бинаризовано*).

- pre_till_pclose — плановое количество дней с даты сбора данных до даты закрытия кредита (бинаризовано*).
- pre_till_fclose — фактическое количество дней с даты сбора данных до даты закрытия кредита (бинаризовано*).
- pre_loans_credit_limit — кредитный лимит (бинаризовано*).
- pre_loans_next_pay_summ — сумма следующего платежа по кредиту (бинаризовано*).
- pre_loans_outstanding — оставшаяся невыплаченная сумма кредита (бинаризовано*).
- pre_loans_total_overdue — текущая просроченная задолженность (бинаризовано*).
- pre_loans_max_overdue_sum — максимальная просроченная задолженность (бинаризовано*).
- pre_loans_credit_cost_rate — полная стоимость кредита (бинаризовано*).
- pre_loans5 — число просрочек до 5 дней (бинаризовано*).
- pre_loans530 — число просрочек от 5 до 30 дней (бинаризовано*).
- pre_loans3060 — число просрочек от 30 до 60 дней (бинаризовано*).
- pre_loans6090 — число просрочек от 60 до 90 дней (бинаризовано*).
- pre_loans90 — число просрочек более чем на 90 дней (бинаризовано*).
- is_zero_loans_5 — флаг: нет просрочек до 5 дней.
- is_zero_loans_530 — флаг: нет просрочек от 5 до 30 дней.
- is_zero_loans_3060 — флаг: нет просрочек от 30 до 60 дней.
- is_zero_loans_6090 — флаг: нет просрочек от 60 до 90 дней.
- is_zero_loans90 — флаг: нет просрочек более чем на 90 дней.
- pre_util — отношение оставшейся невыплаченной суммы кредита к кредитному лимиту (бинаризовано*).
- pre_over2limit — отношение текущей просроченной задолженности к кредитному лимиту (бинаризовано*).

- `pre_maxover2limit` — отношение максимальной просроченной задолженности к кредитному лимиту (бинаризовано*).
- `is_zero_util` — флаг: отношение оставшейся невыплаченной суммы кредита к кредитному лимиту равно 0.
- `is_zero_over2limit` — флаг: отношение текущей просроченной задолженности к кредитному лимиту равно 0.
- `is_zero_maxover2limit` — флаг: отношение максимальной просроченной задолженности к кредитному лимиту равно 0.
- `enc_paym_{0..N}` — статусы ежемесячных платежей за последние N месяцев (закодировано**).
- `enc_loans_account_holder_type` — тип отношения к кредиту (закодировано**).
- `enc_loans_credit_status` — статус кредита (закодировано**).
- `enc_loans_account_cur` — валюта кредита (закодировано**).
- `enc_loans_credit_type` — тип кредита (закодировано**).
- `pclose_flag` — флаг: плановое количество дней с даты открытия кредита до даты закрытия не определено.
- `fclose_flag` — флаг: фактическое количество дней с даты открытия кредита до даты закрытия не определено.

* Область значений поля разбивается на N непересекающихся промежутков. Каждому промежутку случайным образом назначается уникальный номер от 0 до N-1, а значение поля заменяется номером промежутка, которому оно принадлежит.

Варианты работы с данными

В итоговом проекте будет много информации для анализа, так как в настоящих задачах больших компаний приходится иметь дело с большими объёмами данных.

Поскольку их много, они не всегда попадают в оперативную память. Скорее всего, не получится загрузить все датасеты в компьютер с оперативной памятью 4 или 8 Гб.

Поэтому мы предоставим вам скрипт, который позволит обрабатывать данные порциями — по несколько файлов за раз. С помощью этого кода вы сможете обрабатывать данные и на личном компьютере. Если на личном компьютере не хватит ресурсов для обработки данных порциями, используйте Google Colab.

Техническое задание

Что нужно сделать

- 1 Скачайте тренировочный датасет. Распакуйте один файл и посмотрите на его структуру.

Рекомендации по выполнению

Поскольку данных в датасете много (должно получиться порядка 4,5 Гб данных в файлах Parquet, объём которых после распаковки станет в разы больше), предлагается читать его итеративно (то есть по несколько файлов за раз — извлекать нужные фичи и записывать только эти фичи в результирующий датафрейм).

Ознакомьтесь с [кодом для базового сбора признаков и прочтения датасета](#).

Критерии оценки

Код обработки данных запущен и отработан без ошибок. В результате получен пробный датафрейм, состоящий из признаков для обучения модели.

- 2 Подумайте, какие признаки могут быть полезны, и смоделируйте их. Дайте комментарии, почему вы сгенерировали тот или иной признак.

Критерии оценки

Собран итоговый датафрейм, состоящий из признаков для обучения модели.

- 3 После того как вы подготовили датасет с признаками, смёрджите их с целевой переменной и выведите размерность результирующего датафрейма.

Рекомендации по выполнению

Сохраните полученный датафрейм в файл. Если ядро Python зависнет, вы сможете сделать работу заново уже с готовым датасетом.

Критерии оценки

В итоговый датафрейм добавлено значение целевой переменной. Количество записей в итоговом датасете — три миллиона.

- 4 С помощью лучшей модели (той, которая показала наилучшие результаты в исследованиях в предыдущем пункте) сделайте предсказания на тестовом датасете.

Критерии оценки

Выборка поделена на тренировочную и тестовую в отношении 70/30 либо 80/20. Выведена размерность каждой из них.

- 5 Подготовьте автоматизированный пайплайн, который по вызову `fit` будет готовить данные и обучать модель, а по вызову `predict` — делать предсказания на заданном наборе данных. Обучите свой пайплайн подготовки данных и обучения модели и сохраните результат обучения в бинарном формате `pickle`.

Критерии оценки

Пайплайн для подготовки данных и обучения модели написан с помощью `sklearn.pipeline`.

Обученный на всём датасете пайплайн сохранён в виде файла в формате `pickle`.

Формат сдачи материалов

Отправьте на проверку куратору следующие документы:

- Jupyter Notebook со всеми этапами решения задачи.
- Файл с обученным пайплайном подготовки данных и моделью.
- Файл с предиктами на тестовой выборке.

Критерии оценки

В итоговом проекте будет много информации для анализа, так как в настоящих задачах больших компаний приходится иметь дело с большими объёмами данных.

Важно, чтобы были выполнены все пункты заданий, а результаты соответствовали критериям, указанным в ТЗ.

Метрика качества

- Значение метрики на тестовом датасете должно быть не менее 0,75 по ROC-AUC (это минимально допустимое бейслайн-значение, большее значение метрики приветствуется).
- Такое значение должно быть получено не за счёт переобучения на тестовом датасете (при выборе другого тестового сета значение метрики будет схожим).

Чистота кода

- Код соответствует стандарту PEP 8: при проверке особенное внимание будет уделяться комментариям и осмысленным названиям переменных.

Количество и состав проведённых экспериментов

- В рамках экспериментов по feature-инжинирингу и моделированию проведено минимум по три эксперимента для выявления лучших моделей или фичей.
- Подобраны гиперпараметры модели для улучшения финальной метрики.

Проект возвращается на доработку, если:

- Выполнены не все пункты заданий. Результаты по всем или некоторым пунктам заданий не соответствуют критериям, указанным в ТЗ.
- Обученный пайплайн не запускается и не выдаёт предсказания.
- Код не запускается либо запускается, но с ошибками.
- Модель переобучена на тестовом датасете.
- Значение метрики на тестовом датасете — менее 0,75 по ROC-AUC.
- Код нечитаем, плохо декомпозирован.
- Не подтверждена экспериментальная часть итоговой работы.



Презентация итоговой работы

Чтобы получить сертификат о прохождении курса, сдайте проект, дождитесь его одобрения куратором и после этого запишитесь на презентацию в курсе [«Презентация итоговых проектов»](#).

Представление результатов проекта включает устный рассказ по заранее подготовленной презентации и проверку ваших расчётов. Презентация проекта проходит в онлайн-формате: десять минут на презентацию проекта и пять-десять минут на вопросы. Выступление должно состоять из следующих частей:

1

Заглавный слайд с вашим именем и темой работы.

2

Описание проблемы, которую вы решали.

3

Презентация результатов Feature Preparation:

- Feature Engineering (какие из подготовленных фичей попали в итоговую модель по результатам экспериментов).

- Презентация финального датасета (какие из фичей пойдут в модель, описание способов их подготовки для моделирования).

4

Презентация результатов моделирования:

- Сводная таблица апробированных методов с результатами по метрикам на тренировочном (по кросс-валидации) и тестовом датасете.
- Презентация результатов настройки гиперпараметров на лучшей модели (удалось ли достичь улучшения качества за счёт тюнинга гиперпараметров).
- Визуализация итогового графика и значения метрики по ROC-AUC на финальной модели с подобранными гиперпараметрами.

5

Демонстрация конечного результата.

Полезные материалы для выполнения проекта

Стандарты РЕР 8

Материал подготовлен на основе данных соревнования о карточных транзакциях, полученных с сайта Open Data Science (ООО “Соревнования Анализа Данных”) ods.ai.