My first scientific paper

Week 1

Set the toolbox

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Термины:[1]

- модель,
- критерий,
- алгоритм,
- метод.

[1] Синонимия в терминологии — одна из проблем машинного обучения.

Потеря информации при передаче сообщения[2]

Небольшая группа программистов работает над новым проектом. Сколько времени пройдет, прежде чем

1) в группе выработается свой уникальный лабораторный жаргон,
2) новый сотрудник сможет разобраться, чем занимается группа,
3) руководитель группы перестанет понимать ход проекта,
4) каждый член группы перестает понимать, чем занимаются его коллеги?

[2] в отсутствие планирования

Исследователь-аналитик в коммерческой компании

Director

Customer

Analyst

Expert

Исследователь-аналитик в стартапе

Startup team

Investors

Исследователь-аналитик в научной группе

Research team

Fund, company

Исследователь-аналитик в научной группе

Research team

До начала планирования исследования аналитик и (эксперт) обсуждают ключевые вопросы

1. Цель проекта. (Ожидаемый результат разработки.) **Ожидаемая цель исследования.**
2. Прикладная задача, решаемая в проекте. (Как результат будет использован?) **Чем результат будет проиллюстрирован?**
3. Описание исторических измеряемых данных. (Форматы и тайминг.) **Алгебраическая структура данных.**
4. Критерии качества. (Как измеряется качество полученного результата, что будет в отчете?) **Функция ошибки, что будем оптимизировать.**
5. Выполнимость проекта. (Как показать, что проект выполним, список возможных рисков.) **План анализа ошибки.**

Построение скоринговых вероятностных моделей как прикладная задача классификации

- Выдача кредита (Application scoring)
- Динамика состояния (Behavioral scoring)
- Просроченная задолженность (Collection scoring)

Типы кредитов для физических лиц:

- Потребительский (POS)
- Кредит наличными
- Автокредит
- Ипотечный

Типичное число клиентских записей в базе данных:

- $\sim 10^4$ для «тяжелых» долгосрочных кредитов,
- $\sim 10^6$ для «легких» кредитов,
- $\sim 10^7$ для банковских карт.

Процедура получения кредита с точки зрения банка

Заявка-анкета, кредитная история

↓

Скоринговый балл: вероятность невозврата (PD)

Вариант: кредитный контроль

↓

Назначение кредитного лимита

↓

Сообщение о принятии заявки или об отказе

↓

Заключение договора

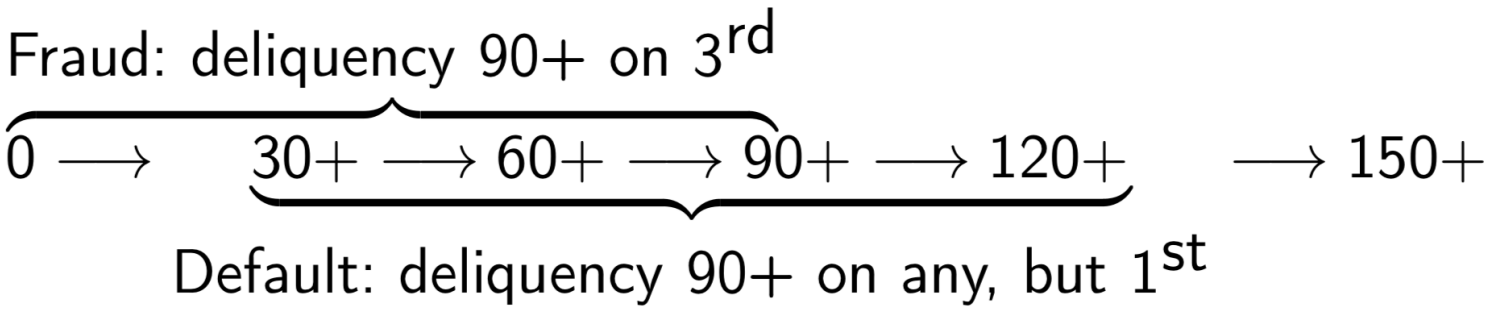
↓

...

↓

Появление кредитной истории

Виды просрочек возврата кредита



- Fraud — мошенничество
- Default — возврат кредита просрочен

Потери от просрочек возврата потребительского кредита

Примерная просрочка (от недели и выше) по потребительским кредитам на некоторый момент времени

Категория	Количество	Сумма
Все категории товаров	100 000	2 100 М
Бытовая техника	30 000	350 М
Мебель	20 000	300 М
Одежда	15 000	200 М
Телевизоры	10 000	100 М
Мобильные телефоны	15 000	80 М
Фотоаппараты	2 000	20 М

Причины отказа в выдаче кредита

Некоторые типичные причины:

- недостаточный скоринговый балл,
- не прошел кредитный контроль,
- в черном списке банка,
- просрочка по данным бюро кредитных историй,
- не гражданин России,
- маленький личный доход,
- клиент моложе (старше) определенного возраста и сумма слишком велика,
- мобильный телефон найден у другого клиента.

Общие сведения о выборке

- Кредиты с просрочкой 90+, дефолты
- Случаи мошенничества (fraud) из выборки исключены
- Всего элементов выборки $\sim 10^4$–10^6
- Доля просрочивших (default rate) ~ 8–16%
- Период наблюдения – не менее 91 дней после заключения контракта
- Число исходных переменных ~ 30–50
- Число пропущенных записей > 0, обычно мало
- Число записей-выбросов > 0, $3\sigma^2$-cutoff

Список переменных

Variable	Type	Categories
Loan currency	Nominal	3
Applied amount	Linear	
Monthly payment	Linear	
Tetm of contract	Linear	
Region of the office	Nominal	7
Day of week of scoring	Linear	
Hour of scoring	Linear	
Age	Linear	
Gender	Nominal	2
Marital status	Nominal	4
Education	Ordinal	5
Number of children	Linear	
Industrial sector	Nominal	27
Salary	Linear	
Place of birth	Nominal	94
...	...	...
Car number shown	Nominal	2

Преобразование шкал

- Область деятельности заемщика, номинальная шкала

Nominal	Tourism	Banking	Education
John	1	0	0
Thomas	0	1	0
Sara	0	0	1

- Образование заемщика, ординальная шкала

Ordinal	Primary	Secondary	Higher
John	1	0	0
Thomas	1	1	0
Sara	1	1	1

До начала планирования исследования аналитик и (эксперт) обсуждают ключевые вопросы

1. Цель проекта. (Ожидаемый результат разработки.) **Ожидаемая цель исследования.**
2. Прикладная задача, решаемая в проекте. (Как результат будет использован?) **Чем результат будет проиллюстрирован?**
3. Описание исторических измеряемых данных. (Форматы и тайминг.) **Алгебраическая структура данных.**
4. Критерии качества. (Как измеряется качество полученного результата, что будет в отчете?) **Функция ошибки, что будем оптимизировать.**
5. Выполнимость проекта. (Как показать, что проект выполним, список возможных рисков.) **План анализа ошибки.**
6. Условия, необходимые для успешного выполнения проекта. (Организация работ.) **Требования к выборке.**
7. Методы решения. (Библиотеки процедур.) **Поставленные гипотезы, оптимальные вероятностные модели.**

To start an *applied* project **an expert** and **an analyst** set

1. Project goal (the expected result of development)
   main purpose of research
2. Project application (how the project result will be applied)
   environment of measures and impacts
3. Historical data description (data formats and timing)
   algebraic structures of data
4. Quality criteria (how the project quality is measured)
   error function
5. Feasibility of the project (how to prove the project feasibility, list possible risks) error analysis

How long the model lives after being put on operation? What replaces it after?

НИР или ОКР? Новизна или технологичность

Эксперт:

(Как долго будет эксплуатироваться модель? Что заменит ее в дальнейшем?)

Аналитик:

Какое влияние окажет исследование на область знаний? Насколько она будет полезна?

Заполните таблицу для описания проекта

Google Docs to shared editing

За какую задачу браться?

1. Масштабность: решение задачи должно влиять на большое число людей, специалистов, лиц принимающих решения.
2. Заброшенность (популярность) задачи. Общая ошибка: решать популярные задачи.
3. Решаемость задачи. Следует выбирать просто и элегантно решаемые задачи.
4. Наша квалификация и готовность к решению: похожие задачи мы уже решали.

Для тех, кто пишет квалификационные работы.

1) Соединяем 2 и 3, решаем
2) кто ты по стилю мышления (алгебраист, геометр, физик, программист, ...),
3) делаем (без избыточных движений) сильную работу.