

My first scientific paper in the field of Machine Learning

Вадим Викторович Стрижов

Московский физико-технический институт

12 февраля 2020 г.

Цель: научиться ставить задачи в области машинного обучения в индустриальных и научных проектах

Семинары включают:

- ▶ теорию выбора моделей и анализа ошибки,
- ▶ разбор примеров постановок задач,
- ▶ обсуждение методов и инструментов планирования при постановке и решении задач.

Семинары **не** включают:

- ▶ разбор отдельных алгоритмов машинного обучения,
- ▶ разбор кода на питоне и других языках,

потому что курсы и материалы по этим темам уже имеются в достаточном количестве.

Подробнее см. <http://www.machinelearning.ru/wiki/index.php?title=M1>

Тематический план

- ▶ Модели и их порождение
- ▶ Функция ошибки и выбор моделей
- ▶ Снижение размерности, порождение признаков
- ▶ Метрическое обучение
- ▶ Обучение представлений и вложения
- ▶ Планирование эксперимента и анализ ошибки

Домашние задания

- ▶ Анкета с вопросами на понимание пройденного материала и на подготовку к следующему занятию
- ▶ Основное задание — постановка задачи, решение задачи, либо вычислительный эксперимент
- ▶ Рецензирование работ — их просмотр и ранжирование

Робот **AutoML** (сейчас в режиме отладки)

1. Собирает проекты из папок GitHub, организация `phystech-intelligent-systems`.
2. Рассылает ссылки на проекты рецензентам (вам, слушателям этого курса).
3. Каждый рецензент получает от трех до пяти ссылок на проекты и ранжирует их по качеству, заполняя анкету (так же, как это делает Курсера).
4. Рассылает оценки за домашнее задание в конце недели.

Потеря информации при передаче сообщения¹

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?

¹ в отсутствие планирования

Исследователь-аналитик в коммерческой компании



Director



Customer



Analyst



Expert

The source: <http://rpelm.com>

Исследователь-аналитик в стартапе



Startup team



Investors

The source:<http://rpelm.com>

Исследователь-аналитик в научной группе



Research team



Fund, company

The source: <http://rpelm.com>

За какую задачу браться?

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

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.**

Построение скоринговых вероятностных моделей как прикладная задача классификации

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

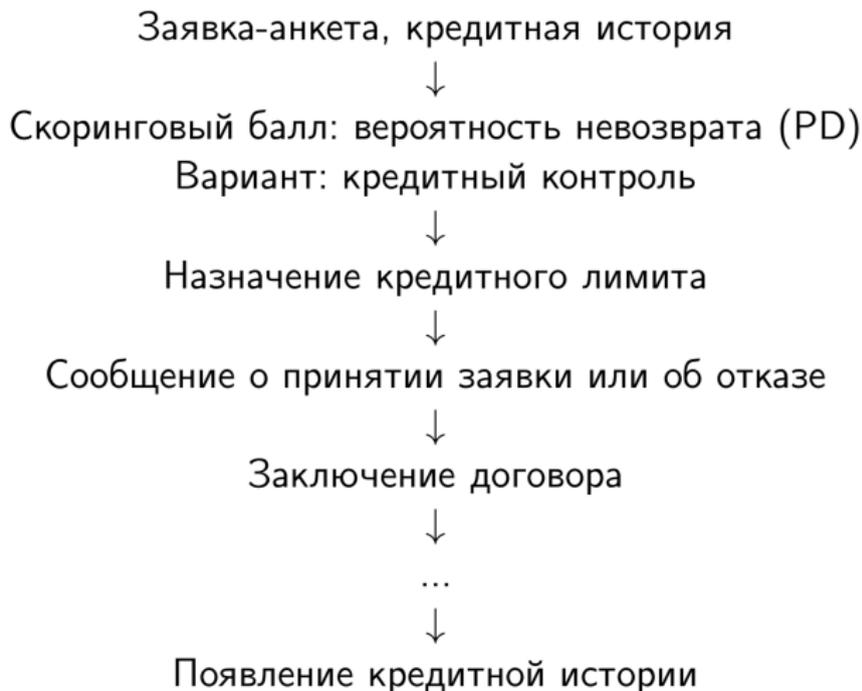
Типы кредитов для физических лиц:

- 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



Виды просрочек возврата кредита

Fraud: delinquency 90+ on 3rd

0 → 30+ → 60+ → 90+ → 120+ → 150+

Default: delinquency 90+ on any, but 1st

- 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

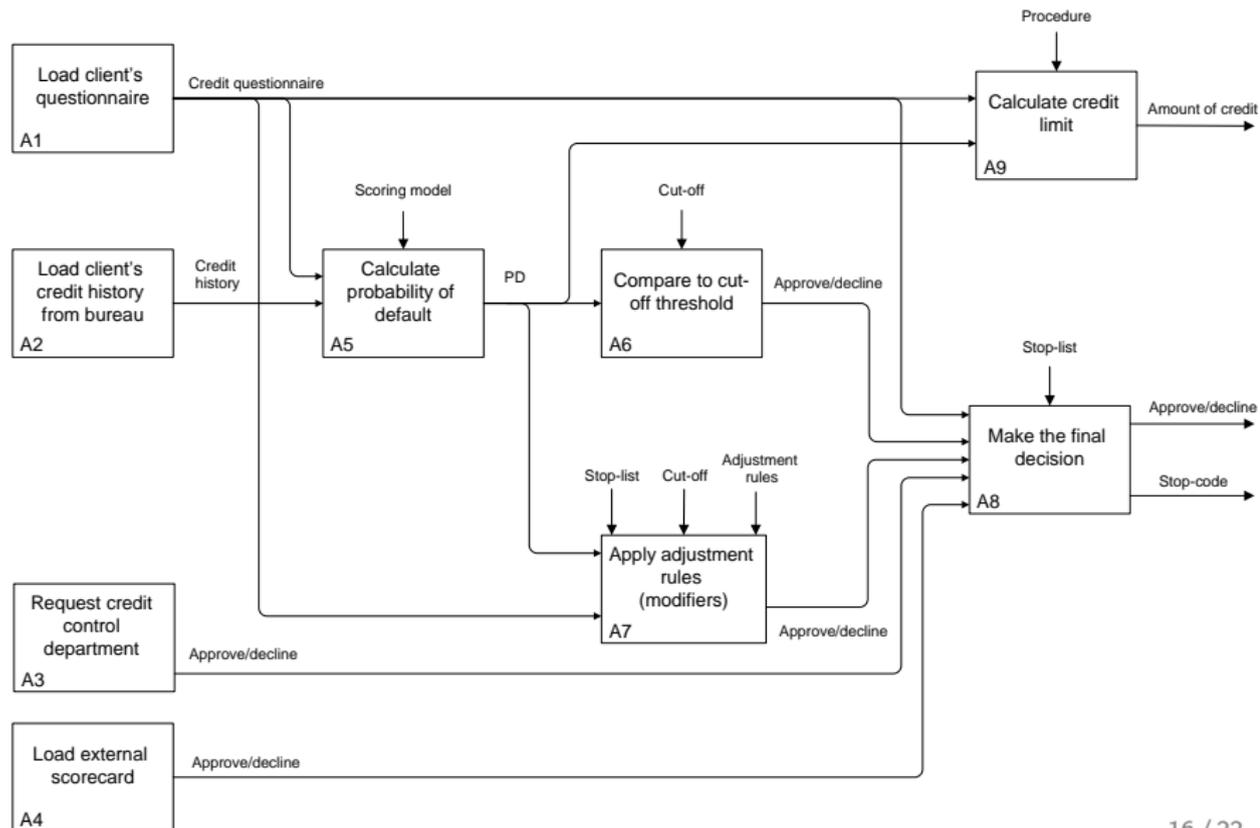
Категория	Количество	Сумма
Все категории товаров	100 000	2 100 М
Бытовая техника	30 000	350 М
Мебель	20 000	300 М
Одежда	15 000	200 М
Телевизоры	10 000	100 М
Мобильные телефоны	15 000	80 М
Фотоаппараты	2 000	20 М

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.

Функциональная схема IDEF0 скоринговой карты с вероятностной моделью классификации



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?**

практика Стрижов - Google S x

Secure | <https://www.google.ru/search?q=практика+Стрижов&oq=практика+Стрижов&aqs=chrome>

Google

практика Стрижов

All Videos Images News Maps More Settings Tools

About 10,400 results (0.39 seconds)

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▼ Translate this page

Jul 7, 2018 - Katrutsa A.M., Strijov V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 ...

Задачи · Домашние задания · Подготовка к экзамену · Справочные материалы

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▼ Translate this page

Aug 24, 2017 - Карасиков М.Е., Стрижов В.В. Классификация временных рядов в пространстве параметров порождающих моделей // Информатика и ...

(практика, В.В. Стрижов)/Группа 474, весна 2018 - MachineLearning.ru

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▼ Translate this page

Jun 20, 2018 - Короткая ссылка bit.ly/2JkLqlo. Два курса весеннего семестра. Выбор моделей в задачах регрессии и классификации, лекции по теории ...

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▼ Translate this page

May 23, 2018 - Адуенко А.А. Выбор мультимоделей в задачах классификации (научный руководитель В.В. Стрижов). Московский физико-технический ...