Как не работать в стол: управление проектами с высокой ценой ошибки

Алексей Арустамов

Loginom Company (ex. BaseGroup Labs)

Практическое применение машинного обучения

- Поиск
- Рекомендации
- Обогащение данных
- Анализ изображений
- Переводы
- Игры

Доля отраслей в ВВП России*

По данным Росстата за 2016 год

Задачи, с максимальным экономическим эффектом:

- Оптимизация запасов
- Управление рисками
- Производство
- Логистика
- Клиентская лояльность

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

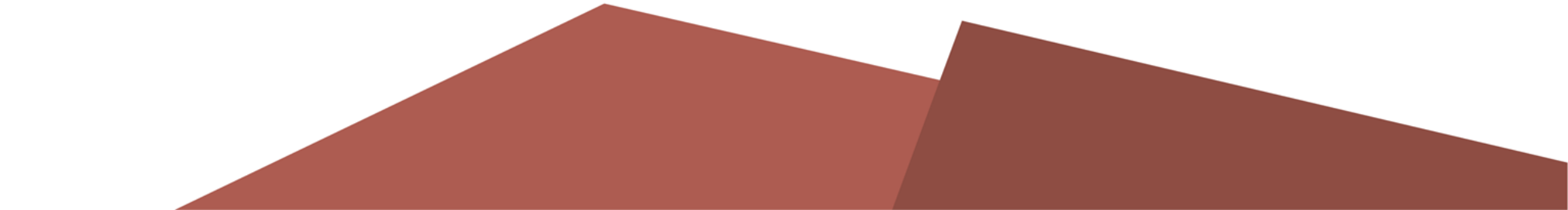
Григорий Пятецкий-Шапиро

Высокая цена ошибки предполагает высокое доверие к системе

Основа доверия – предсказуемость:

1. Осознание ограничений
2. Понимание всех шагов
3. Аргументация для каждого шага
4. Тестирование
5. Поведение в нестандартных случаях

Кейс: Кредитный скоринг



Эволюция скоринга:

1. Жесткие правила
2. Дженерик модели
3. Индивидуальные модели
4. Промышленное моделирование

Паспорт скоринговой модели – основной инструмент доказательства её адекватности.

Сфера применения:

- Параметры продукта
- Диапазон сумм
- Диапазон сроков
- Время построения
- Регион
- Дополнительные ограничения

Подготовка: Первичный аудит

Статистика:

- Пропуски
- Выбросы
- Аномалии
- Дубли

Бизнес:

- Недостоверные данные
- Несогласующиеся данные

Подготовка: Исключаемые атрибуты

- Плохая статистика
- Данные из будущего
- Несоответствие бизнес-требованиям

Объяснения для каждого атрибута.

Подготовка: Редкие значения

Правила обработки каждого атрибута:

- Исключить записи
- Работать как с пропусками
- Заменять по оговоренному правилу

Подготовка: Порождение атрибутов

Описание формул и правил формирования каждого атрибута:

- Производные характеристики
- Кросс-характеристики

Подготовка: Обязательные атрибуты

- Требования документооборота
- Требования законодательства
- Требования бизнес-процессов
- Значимые с точки зрения рисковиков

Подготовка: Определение события

Описание, что мы считаем целевой переменной, с аргументацией почему выбрана именно она.

Подготовка: Итоговая выборка

- Правила формирования обучающего и тестового множества
- Статистика по всему набору, обучающему и тестовому множествам
- Распределение в разрезе целевой переменной

Отбор атрибутов

- Корреляционный анализ
- Оценка значимости WoE
- Оценка предсказательной силы IV

Моделирование: Построение

Начинать с логистической регрессии:

- Знакома рисковикам
- Легко интерпретируется
- Является базой для сравнения

Описание логики расчета балла отсечения.

Оценка качества

- Индекс Gini
- ROC-кривая
- Статистика Колмогорова-Смирнов
- Распределение плохой-хороший
- Диаграмма изменения уровней одобрения

Правила мониторинга

- Регулярность мониторинга
- Правила расчета индикаторов
- Логика оповещения

Финальная проверка

Построенная модель не должна противоречить экономическому смыслу и вызывать отторжения экспертов. Явный негатив резко снижает доверие к выводам.

Что повышает доверие:

1. Указание области применения и ограничений
2. Объяснение всех преобразований
3. Правила обработки в граничных случаях
4. Объяснение отбора факторов
5. Объяснение модели, если возможно
6. Тестирование модели
7. Расчет индикаторов качества моделей
8. Осмысленность с точки зрения бизнес-логики

Доверие необходимо взращивать от простых моделей к более сложным, тогда цена ошибки не будет стоп-фактором.

Самообучающийся чат-бот от Microsoft за 24 часа превратился из мизантропа в нациста

loginom.ru