Скоринговые карты - руководство пользователя

Содержание

Оглавление

Глоссарий

Биннинг	Процесс разбиения непрерывной числовой переменной на интервалы (группы), называемые бинами, для последующего анализа и преобразования.
WOE (Weight of Evidence, Вес доказательств)	Статистическая мера, показывающая, насколько вероятность принадлежности к целевому классу (например, дефолту) отличается внутри конкретной группы (бина) от общей вероятности в выборке. Используется для кодирования и оценки информативности признаков.
IV (Information Value, Информационная ценность)	Суммарный показатель прогностической силы признака, рассчитанный на основе WOE. Используется для отбора наиболее значимых для модели переменных. Признаки с низким IV, как правило, исключаются.
Целевая переменная (target)	Зависимая переменная, которую модель обучается предсказывать. В контексте кредитного скоринга — это бинарная переменная, где 1 обычно обозначает "плохого" заемщика (дефолт), а 0 — "хорошего".
EDA (Exploratory Data Analysis, Разведывательный анализ данных)	Начальный этап анализа данных, направленный на выявление их структуры, закономерностей, аномалий и проверку качества (пропуски, выбросы, распределения).
Мультиколлинеарность	Ситуация, когда две или более независимых переменных (признаков) в модели сильно коррелируют между собой. Это может искажать оценки модели и снижать её интерпретируемость.
Порог корреляции Пирсона	Заданное пользователем значение коэффициента корреляции (например, 0.7), при превышении которого для пары признаков система исключает один из них для устранения мультиколлинеарности.
Редкие категориальные значения	Категории в категориальном признаке, частота встречаемости которых в данных ниже заданного порога (MIN_THRESHOLD). Такие значения обычно объединяются в общую группу (например, "Прочее").
Балансировка классов	Техника обработки данных для устранения дисбаланса в распределении целевой переменной (например, когда "плохих" заемщиков значительно меньше, чем "хороших"). В системе применяется метод RandomOverSampler (случайное дублирование примеров миноритарного класса).
ROC-AUC (Area Under Curve)	Площадь под ROC-кривой (Receiver Operating Characteristic). Основная метрика качества бинарного классификатора, показывающая его способность разделять классы. Значение от 0.5 (случайное угадывание) до 1.0 (идеальное разделение).
Коэффициент Джини (Gini)	Производная метрика от AUC (Gini = 2 * AUC - 1). Широко используется в финансовой аналитике и скоринге для оценки разделяющей способности модели.
KS-статистика (Kolmogorov-Smirnov)	Максимальное расстояние между кумулятивными функциями распределения баллов для "хороших" и "плохих" заемщиков. Показывает, насколько хорошо модель различает классы.

PSI (Population Stability Index, Индекс стабильности популяции)	Метрика, используемая для сравнения распределения скоринговых баллов в двух выборках (например, обучающей и текущей). Позволяет оценить стабильность модели во времени.
Скоринговая карта (Scorecard)	Финальная таблица, в которой каждому возможному значению или интервалу признака (бина) присвоено определённое количество баллов. Сумма баллов по всем признакам даёт итоговый скоринговый балл клиента.
Пайплайн (Pipeline)	Автоматизированная последовательность шагов обработки данных и построения модели, от загрузки сырых данных до генерации скоринговой карты и отчётов.
OneHotEncoder	Метод преобразования категориальных признаков в бинарный (0/1) формат, где каждая уникальная категория становится отдельным столбцом.
StandardScaler	Алгоритм стандартизации, который преобразует числовые признаки таким образом, чтобы их среднее значение было равно 0, а стандартное отклонение — 1.
Логистическая регрессия	Статистическая модель, используемая для решения задач бинарной классификации. В контексте скоринга часто применяется благодаря своей интерпретируемости и возможности получения вероятностей.
Артефакты модели	Файлы, генерируемые системой в процессе и по итогам работы: обученная модель, скоринговая карта, таблицы с преобразованными данными, графики и отчёты.
ID (Идентификатор)	Уникальный номер, присвоенный каждому клиенту или записи в наборе данных. Используется для однозначной идентификации наблюдений.
Портфель (Portfolio)	Название набора данных или конкретного кредитного продукта, для которого строится скоринговая карта (задаётся параметром PORTFOLIO).

1 Назначение системы

Scorecard Modeler — комплексное решение, автоматизирующее процесс построения скоринговых карт для оценки рисков, связанных с клиентами. Данное решение позволяет на основе анализа сотен характеристик заемщика количественно оценить риски и предсказать вероятность возврата кредита, что существенно повышает точность и эффективность принятия решений в финансовых учреждениях.

1.1 Основные характеристики и преимущества

- **Автоматизация процесса**: Скоринг-моделирование с использованием автоматизированных алгоритмов позволяет значительно сократить время на создание и настройку скоринговых карт, а также минимизировать количество ошибок, связанных с ручным анализом данных.
- **Многообразие характеристик заемщика**: Модель учитывает различные факторы, включая финансовую историю, поведение клиента, кредитные и демографические данные, что позволяет создать более точную и надежную модель риска.
- **Количественная оценка рисков**: Благодаря использованию математических и статистических методов, Scorecard Modeler помогает количественно оценить вероятность дефолта заемщика, предоставляя объективную основу для принятия кредитных решений.
- **Прогнозирование вероятности возврата кредита**: Модель позволяет предсказать вероятность того, что заемщик вернет кредит, что помогает банкам и микрофинансовым организациям принимать более обоснованные решения при одобрении или отклонении заявок.
- **Создание балльных скоринговых карт**: С помощью решения можно эффективно создавать скоринговые карты, которые могут быть использованы для быстрой и точной оценки рисков на основе множества входных данных.
- **Область применения**: Scorecard Modeler применяется для создания балльных скоринговых карт в различных финансовых учреждениях, таких как банки, микрофинансовые компании, страховые компании и другие организации, работающие с кредитованием и управлением рисками.

Преимущества:

- **Увеличение точности**: Обеспечивает более точное и надежное прогнозирование вероятности возврата кредита на основе аналитики больших данных.
- **Снижение рисков**: Помогает минимизировать финансовые потери, связанные с непогашением кредитов.
- **Эффективность**: Ускоряет процесс принятия решений, улучшая операционные процессы и позволяя оперативно реагировать на изменения в рисках.

1.2 Область применения

- Банки и микрофинансовые организации
- Кредитные кооперативы

- Страховые компании
- Финтех-компании

2 Роли и доступ

Пользователь:

- Полный доступ ко всем функциям системы
- Настройка параметров моделирования
- Управление пользователями
- Мониторинг выполнения задач
- Загрузка и подготовка данных
- Настройка параметров обработки
- Запуск и мониторинг процессов построения модели
- Анализ результатов и метрик
- Просмотр готовых скоринговых карт
- Анализ отчетов и метрик качества
- Экспорт результатов

3 Навигация по интерфейсу

3.1 Интерфейс приложения

Интерфейс реализован на библиотеке tkinter и позволяет настраивать ключевые параметры обработки данных без редактирования кода.
На главном меню отображается интерфейс мастер настроек:

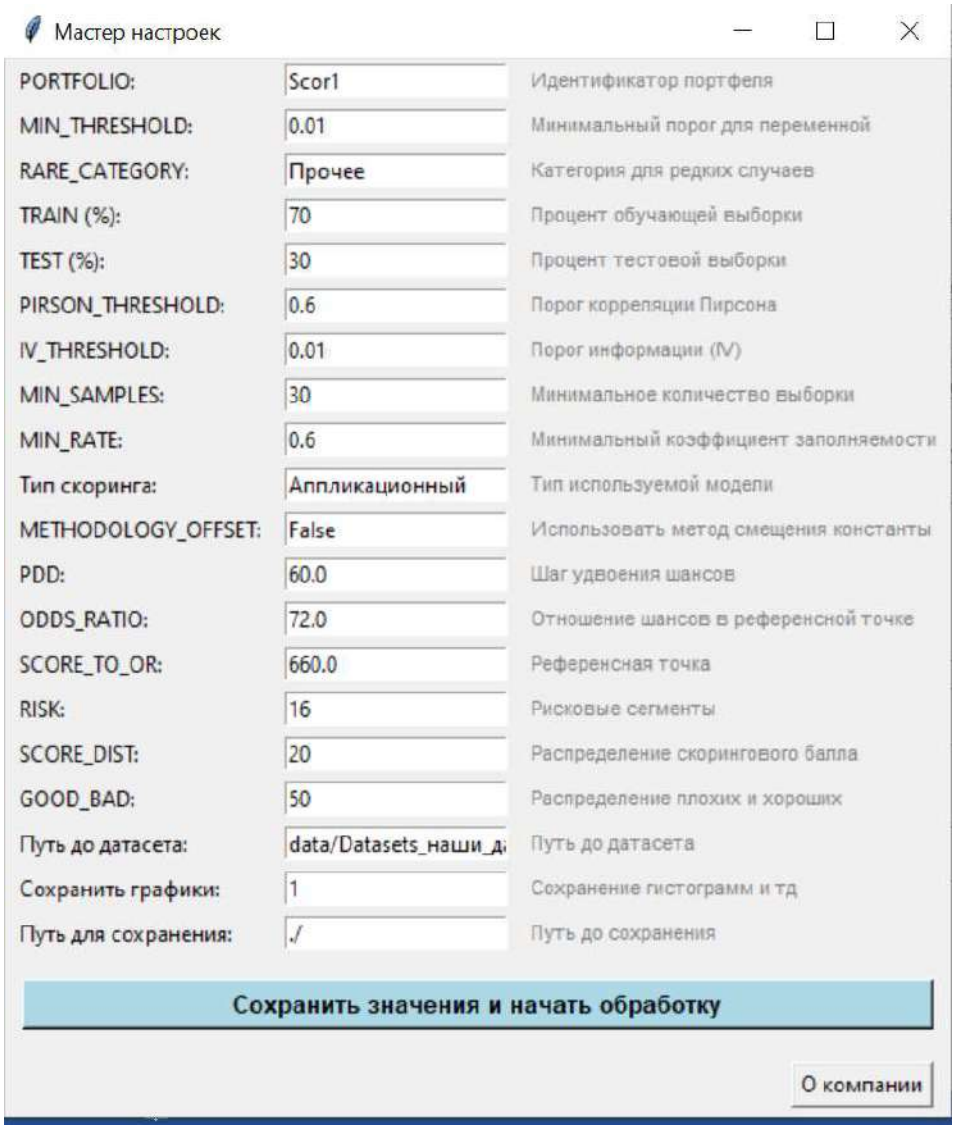

Рис. 1: Структура главного меню Scorecard Modeler

3.2 Этапы работы программы

1. **Мастер настройки**: Конфигурация параметров обработки
2. **Загрузка данных**: Импорт исходных данных
3. **EDA**: Разведывательный анализ данных
4. **Предобработка**: Очистка и подготовка данных
5. **Моделирование**: Построение и оценка моделей
6. **Отчеты**: Генерация скоринговых карт и отчетов
7. **Администрирование**: Управление системой

4 Мастер настройки параметров

4.1 Общие параметры

Параметр	Описание	Тип/Значение
PORTFOLIO	Название портфеля/проекта	Строка (Scor1)
MIN_THRESHOLD	Минимальный порог частоты для редких категорий	Число (0.01)
RARE_CATEGORY	Наименование для редких категорий	Строка ("Прочее")
TRAIN	Доля обучающей выборки (%)	Целое число (70)
TEST	Доля тестовой выборки (%)	Целое число (30)
PIRSON_THRESHOLD	Порог корреляции Пирсона	Число (0.6)
NUM_BINS	Количество бинов для биннинга	Целое число (5)
IV_THRESHOLD	Порог информационной ценности (IV)	Число (0.01)
MIN_SAMPLES	Минимум наблюдений в переменной	Целое число (30)
MIN_RATE	Минимальная доля заполненности признака	Число (0.6)

Таблица 1: Параметры мастера настройки

5 Загрузка исходных данных

5.1 Требования к данным

- **Форматы**: CSV, Excel (.xlsx, .xls)
- **Обязательные поля**:
  - ID — уникальный идентификатор клиента
  - target — целевая переменная (0 - "хороший 1 - "плохой")
- **Кодировка**: UTF-8
- **Разделитель**: Запятая или точка с запятой

5.2 Результат работы программы:

В результате полного цикла обработки данных и построения скоринговой модели в Scorecard Modeler формируется набор входных файлов, предназначенных для анализа, валидации и последующего внедрения скоринговой модели

- Result –каталог/или логический раздел, содержащий все выходные артефакты, сформированные в ходе выполнения скорингового пайплайна.
- Данные для обучения - файл, содержащий итоговый датасет, использованный для обучения скоринговой модели.

  Включает:

  • исходные атрибуты после очистки данных;

• отобранные признаки, прошедшие этапы корреляционного анализа и отбора;

• целевую переменную (target);

• идентификатор объекта (ID).

Файл используется для:

• воспроизводимости обучения модели;

• повторной валидации;

• анализа качества данных.

- Конечные классы категориальные - файл, содержащий результат биннинга категориальных признаков.

  Для каждого категориального атрибута указываются:

  • исходные категории;

  • объединённые классы (бины);

  • соответствующие значения WOE (Weight of Evidence);

  • показатели информативности (IV).

  Файл используется для:

  • интерпретации модели;

  • анализа вклада категориальных признаков;

  • дальнейшего внедрения модели в продуктивные системы.

- Конечные классы непрерывные - файл, содержащий результат биннинга непрерывных (числовых) признаков.

  Для каждого непрерывного атрибута указываются:

  • интервалы значений;

  • границы бинов;

  • значения WOE;

  • вклад интервалов в итоговый скоринг.

  Файл отражает логику разбиения числовых признаков и используется для:

  • прозрачности модели;

  • проверки корректности биннинга;

  • переноса модели в другие системы.

- Преобразованные атрибуты - файл, содержащий значения признаков после применения всех преобразований, включая:

  • биннинг;

  • замену значений на WOE;

  • исключение неинформативных атрибутов.

  Файл представляет собой финальный набор признаков, используемых непосредственно в расчёте скорингового балла.

- Распределение Score по ID - файл, содержащий рассчитанное скоринговое значение (Score) для каждого объекта с уникальным идентификатором (ID).

  Для каждой записи указывается:

  • ID объекта;

  • рассчитанный скоринговый балл;

  • при необходимости — класс риска или сегмент.

  Файл используется для:

  • анализа распределения скоринга;

  • оценки качества сегментации;

  • визуализации и бизнес-аналитики.

- Скоринговая карта - итоговый файл, содержащий формализованное описание скоринговой модели.

  Включает:

  • список используемых признаков;

  • бины и соответствующие баллы;

  • коэффициенты модели;

  • правила расчёта итогового Score.

  Скоринговая карта является основным артефактом для:

  • промышленного внедрения модели;

  • передачи в ИТ-системы заказчика;

  • сопровождения и аудита модели.

5.3 Формат скоринговой карты

Признак	Бин	WOE	Коэффициент	Баллы
Возраст	18-25	0.45	0.23	15
Возраст	26-35	0.12	0.06	4
Возраст	36-50	-0.08	-0.04	-3
Возраст	51+	-0.35	-0.18	-12
Доход	<30k	-0.78	-0.40	-26
Доход	30k-60k	0.23	0.12	8
Доход	60k-100k	0.45	0.23	15
Доход	> 100k	0.65	0.33	22
Стаж	<1 год	-0.65	-0.33	-21
Стаж	1-3 года	0.05	0.03	2
Стаж	3-5 лет	0.25	0.13	9
Стаж	> 5 лет	0.48	0.25	17

Таблица 5: Пример фрагмента скоринговой карты

6 Типовые сценарии работы

6.1 Создание новой скоринговой карты

1. Настройка параметров в мастере
2. Загрузка исторических данных
3. Анализ метрик качества
4. Экспорт скоринговой карты

6.2 Мониторинг качества модели

- Ежемесячный расчет PSI
- Контроль метрик на валидационной выборке
- Анализ распределения баллов
- Проверка стабильности признаков

7 Решение проблем

7.1 Частые проблемы и решения

Проблема	Решение
Низкое значение AUC/Gini	Проверить качество данных, добавить признаки, настроить параметры модели
Высокая корреляция признаков	Увеличить PIRSON_THRESHOLD, удалить коррелирующие признаки
Дисбаланс классов	Использовать балансировку, настроить веса классов
Переобучение модели	Увеличить регуляризацию, добавить больше данных, использовать кросс-валидацию
Долгая обработка данных	Использовать более мощное железо

8 Технические требования

8.1 Аппаратные требования

Компонент	Рекомендации
Процессор	4+ ядер, 1+ GHz
Оперативная память	4+ GB
Дисковое пространство	500+ Mb свободного места

8.2 Программные требования

- Операционная система: Windows 10+