Loginom 6.2 - технический обзор нововведений

Алексей Субботин

Loginom Company

Новое в версиях 6.1-6.2

1. Востребованные алгоритмы
2. Удобное проектирование
3. Легкость интерпретации

Алгоритмы

Регрессия и бинарная классификация

Логистическая и линейная регрессия – самые популярные методы:

- Алгоритмы первого выбора
- Широкая применимость
- Понятная интерпретация

Логистическая и линейная регрессия

Автоматическая настройка для «прикидочного» моделирования с выбором приоритета:

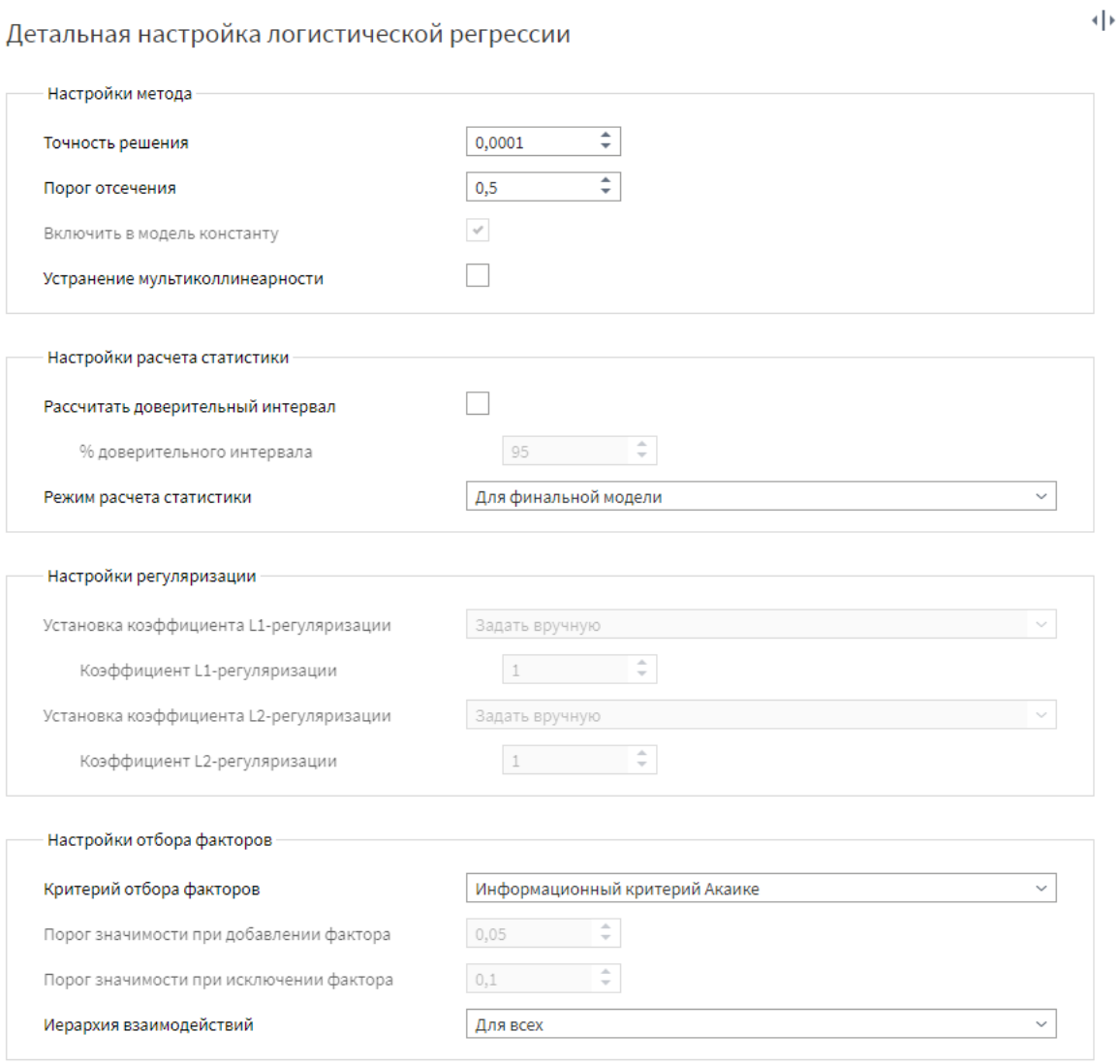
- Максимальная скорость
- Повышенная скорость
- Средние скорость/точность
- Повышенная точность
- Максимальная точность

Логистическая и линейная регрессия

Отбор факторов и построение модели с выбором приоритетов достоверные/недостоверные данные, больше/меньше факторов, точность/скорость:

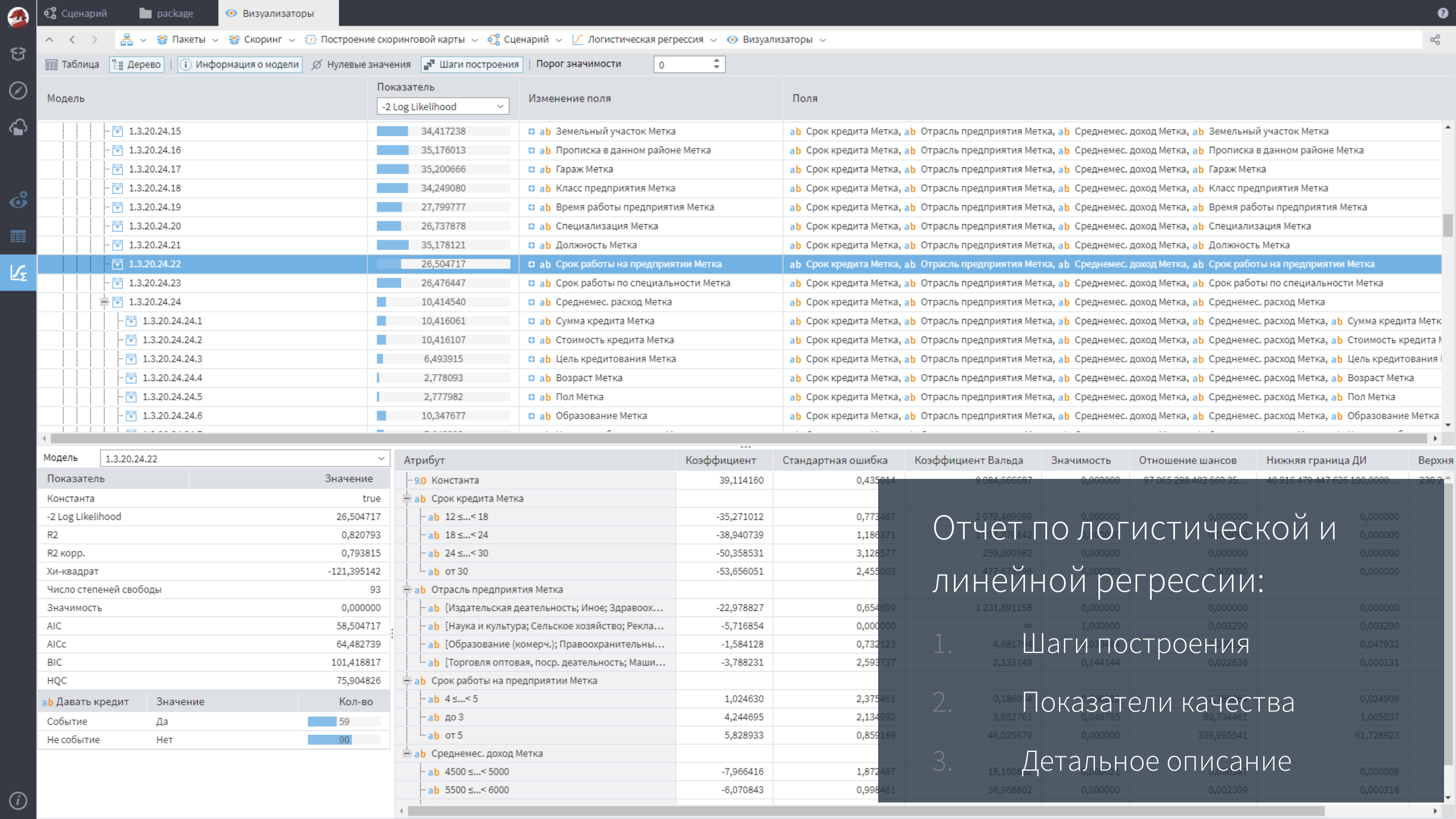
- Классические алгоритмы: Enter, Forward, Backward, Stepwise
- Современные алгоритмы с защитой от переобучения: LASSO, Ridge, Elastic-Net

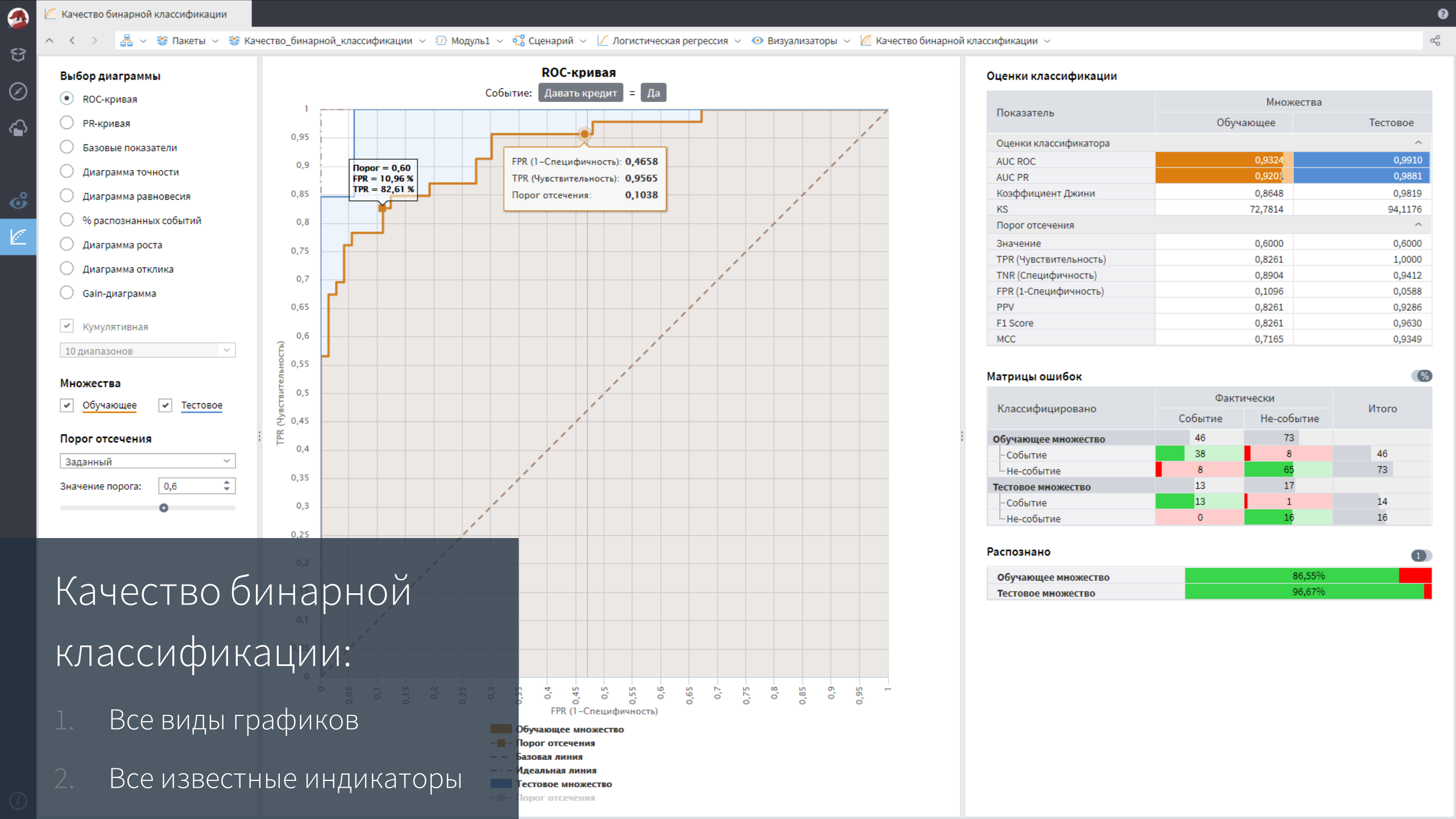
Логистическая и линейная регрессия – экспертный режим для тонкой настройки моделей

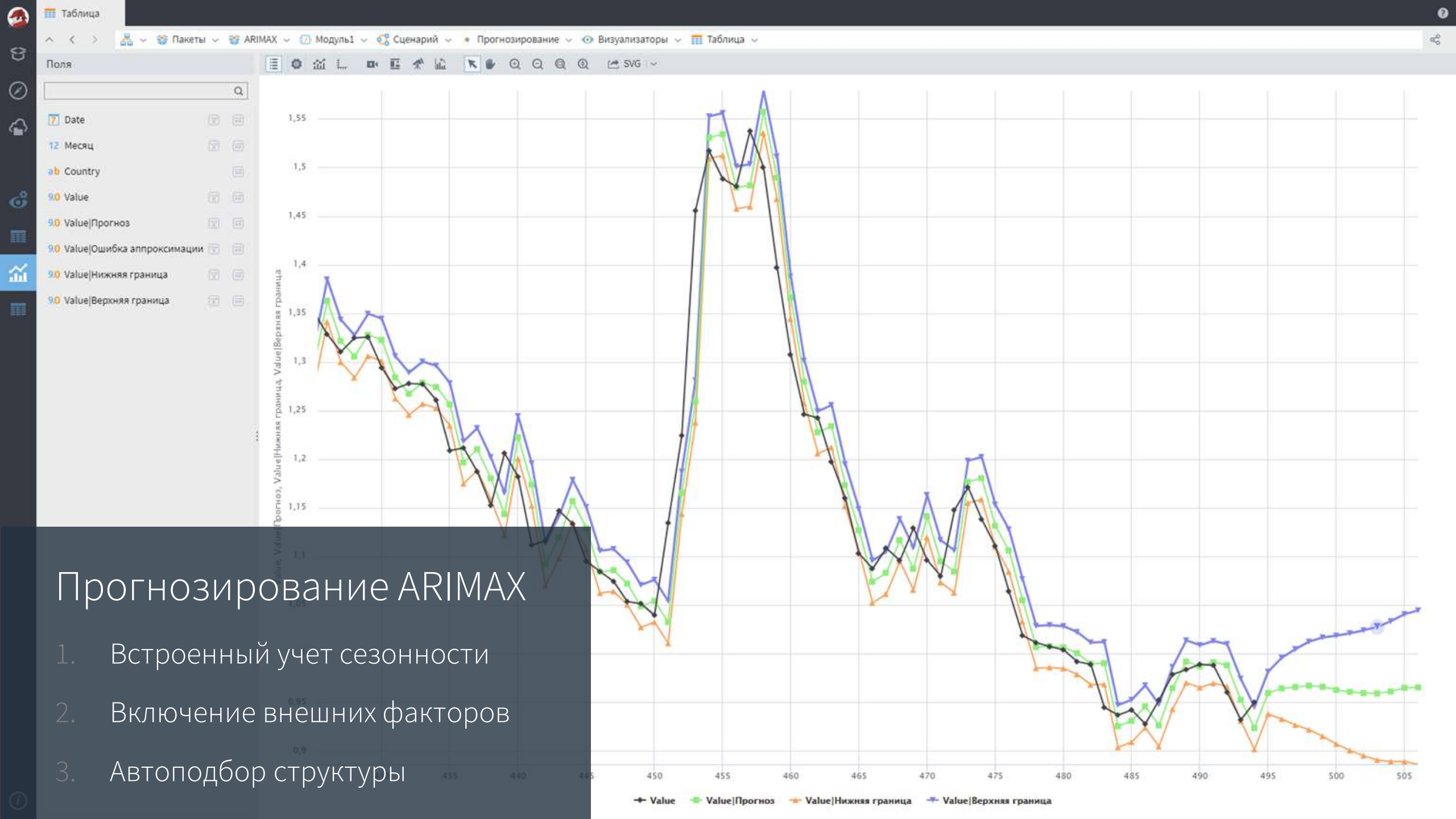


Логистическая и линейная регрессия

1. Построение модели в пару кликов
2. Отбор факторов и подбор лучшей модели
3. Тонкая настройка для экспертов

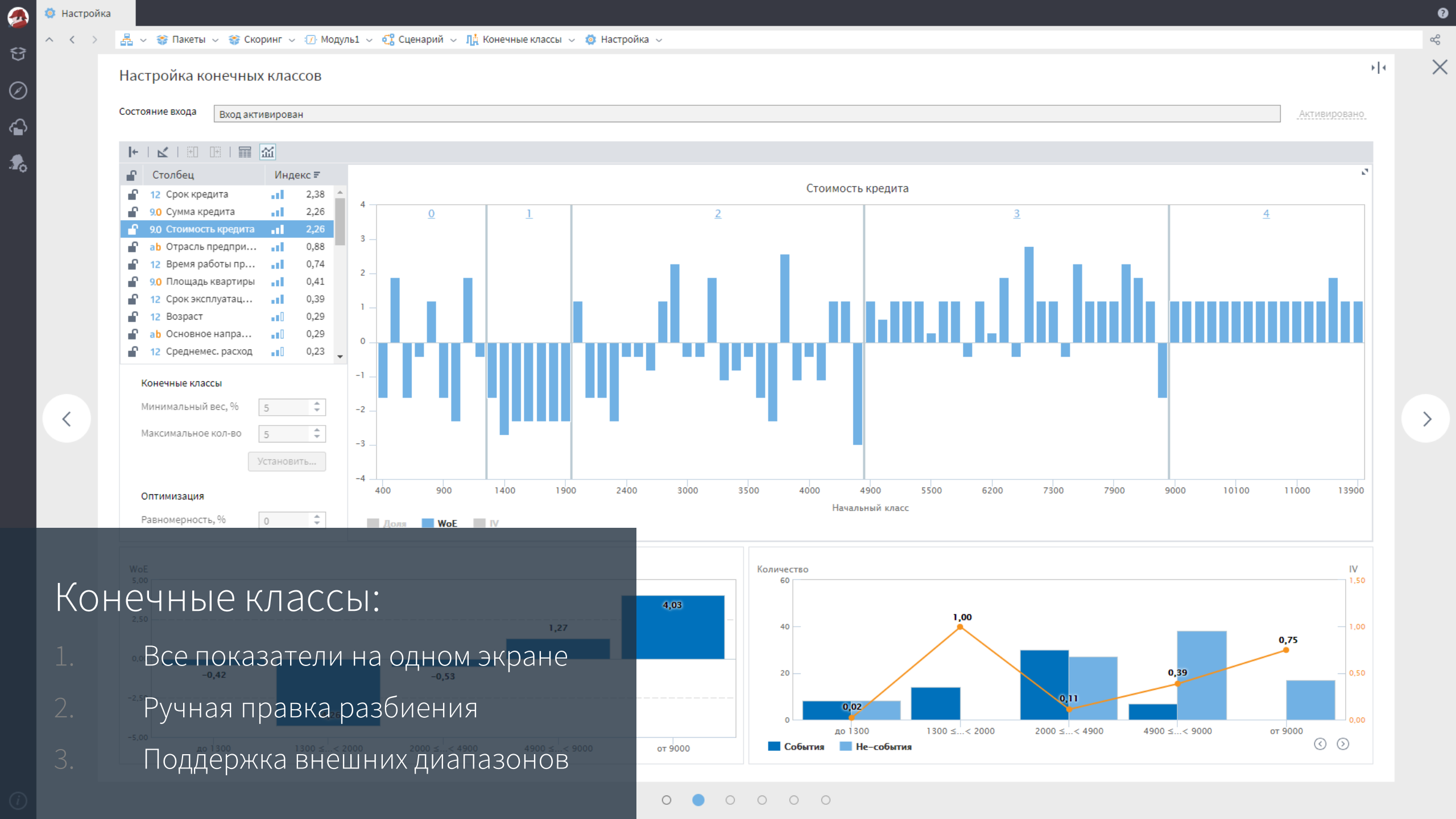

Сценарий
package
Визуализаторы
Пакеты
Скоринг
Построение скоринговой карты
Сценарий
Логистическая регрессия
Визуализаторы
Таблица
Дерево
Информация о модели
Нулевые значения
Шаги построения
Порог значимости
0
Модель
Показатель
-2 Log Likelihood
Изменение поля
Поля
1.3.20.24.15 34,417238 Земельный участок Метка
1.3.20.24.16 35,176013 Прописка в данном районе Метка
1.3.20.24.17 35,200666 Гараж Метка
1.3.20.24.18 34,249080 Класс предприятия Метка
1.3.20.24.19 27,799777 Время работы предприятия Метка
1.3.20.24.20 26,737878 Специализация Метка
1.3.20.24.21 35,178121 Должность Метка
1.3.20.24.22 26,504717 Срок работы на предприятии Метка
1.3.20.24.23 26,476447 Срок работы по специальности Метка
1.3.20.24.24 10,414540 Среднемес. расход Метка
1.3.20.24.24.1 10,416061 Сумма кредита Метка
1.3.20.24.24.2 10,416107 Стоимость кредита Метка
1.3.20.24.24.3 6,493915 Цель кредитования Метка
1.3.20.24.24.4 2,778093 Возраст Метка
1.3.20.24.24.5 2,777982 Пол Метка
1.3.20.24.24.6 10,347677 Образование Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Земельный участок Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Прописка в данном районе Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Гараж Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Класс предприятия Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Время работы предприятия Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Специализация Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Должность Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Срок работы на предприятии Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Срок работы по специальности Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Сумма кредита Метк
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Стоимость кредита М
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Цель кредитования
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Возраст Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Пол Метка
Срок кредита Метка, Отрасль предприятия Метка, Среднемес. доход Метка, Среднемес. расход Метка, Образование Метка
Модель 1.3.20.24.22
Показатель Значение
Константа true
-2 Log Likelihood 26,504717
R2 0,820793
R2 корр. 0,793815
Хи-квадрат -121,395142
Число степеней свободы 93
Значимость 0,000000
AIC 58,504717
AICc 64,482739
BIC 101,418817
HQC 75,904826
Давать кредит Значение Кол-во
Событие Да 59
Не событие Нет 90
Атрибут Коэффициент Стандартная ошибка Коэффициент Вальда Значимость Отношение шансов Нижняя граница ДИ Верхня
Константа 39,114160 0,435014
Срок кредита Метка
12 ≤...< 18 -35,271012 0,773467 2 079,469099 0,000000 0,000000 0,000000
18 ≤...< 24 -38,940739 1,186371 1 077,842… 0,000000 0,000000 0,000000
24 ≤...< 30 -50,358531 3,128577 259,090982 0,000000 0,000000 0,000000
от 30 -53,656051 2,455903 477,676… 0,000000 0,000000 0,000000
Отрасль предприятия Метка
[Издательская деятельность; Иное; Здравоох… -22,978827 0,654699 1 231,891158 0,000000 0,000000 0,000000
[Наука и культура; Сельское хозяйство; Рекла… -5,716854 0,000000 ∞ 1,000000 0,003290 0,003290
[Образование (комерч.); Правоохранительны… -1,584128 0,732123 4,681… 0,030… 0,047932
[Торговля оптовая, поср. деятельность; Маши… -3,788231 2,593737 2,133149 0,144144 0,022636 0,000131
Срок работы на предприятии Метка
4 ≤...< 5 1,024630 2,375461 0,186… 0,024908
до 3 4,244695 2,134992 3,952761 0,046795 69,734461 1,005037
от 5 5,828933 0,859189 46,025679 0,000000 339,995541 61,728923
Среднемес. доход Метка
4500 ≤...< 5000 -7,966416 1,872467 18,100… 0,000021 0,000347 0,000008
5500 ≤...< 6000 -6,070843 0,998461 36,968802 0,000000 0,002309 0,000318
Отчет по логистической и линейной регрессии:
1. Шаги построения
2. Показатели качества
3. Детальное описание


Качество бинарной классификации
Пакеты
Качество_бинарной_классификации
Модуль1
Сценарий
Логистическая регрессия
Визуализаторы
Качество бинарной классификации
Выбор диаграммы
ROC-кривая
PR-кривая
Базовые показатели
Диаграмма точности
Диаграмма равновесия
% распознанных событий
Диаграмма роста
Диаграмма отклика
Gain-диаграмма
Кумулятивная
10 диапазонов
Множества
Обучающее
Тестовое
Порог отсечения
Заданный
Значение порога: 0,6
ROC-кривая
Событие: Давать кредит = Да
Порог = 0,60
FPR = 10,96 %
TPR = 82,61 %
FPR (1–Специфичность): 0,4658
TPR (Чувствительность): 0,9565
Порог отсечения: 0,1038
TPR (Чувствительность)
FPR (1–Специфичность)
Обучающее множество
Порог отсечения
Базовая линия
Идеальная линия
Тестовое множество
Порог отсечения
Оценки классификации
Показатель
Множества
Обучающее
Тестовое
Оценки классификатора
AUC ROC 0,9324 0,9910
AUC PR 0,9201 0,9881
Коэффициент Джини 0,8648 0,9819
KS 72,7814 94,1176
Порог отсечения
Значение 0,6000 0,6000
TPR (Чувствительность) 0,8261 1,0000
TNR (Специфичность) 0,8904 0,9412
FPR (1-Специфичность) 0,1096 0,0588
PPV 0,8261 0,9286
F1 Score 0,8261 0,9630
MCC 0,7165 0,9349
Матрицы ошибок
%
Классифицировано
Фактически
Событие
Не-событие
Итого
Обучающее множество 46 73
Событие 38 8 46
Не-событие 8 65 73
Тестовое множество 13 17
Событие 13 1 14
Не-событие 0 16 16
Распознано
1
Обучающее множество 86,55%
Тестовое множество 96,67%
Качество бинарной классификации:
1. Все виды графиков
2. Все известные индикаторы



Прогнозирование ARIMAX

1. Встроенный учет сезонности
2. Включение внешних факторов
3. Автоподбор структуры



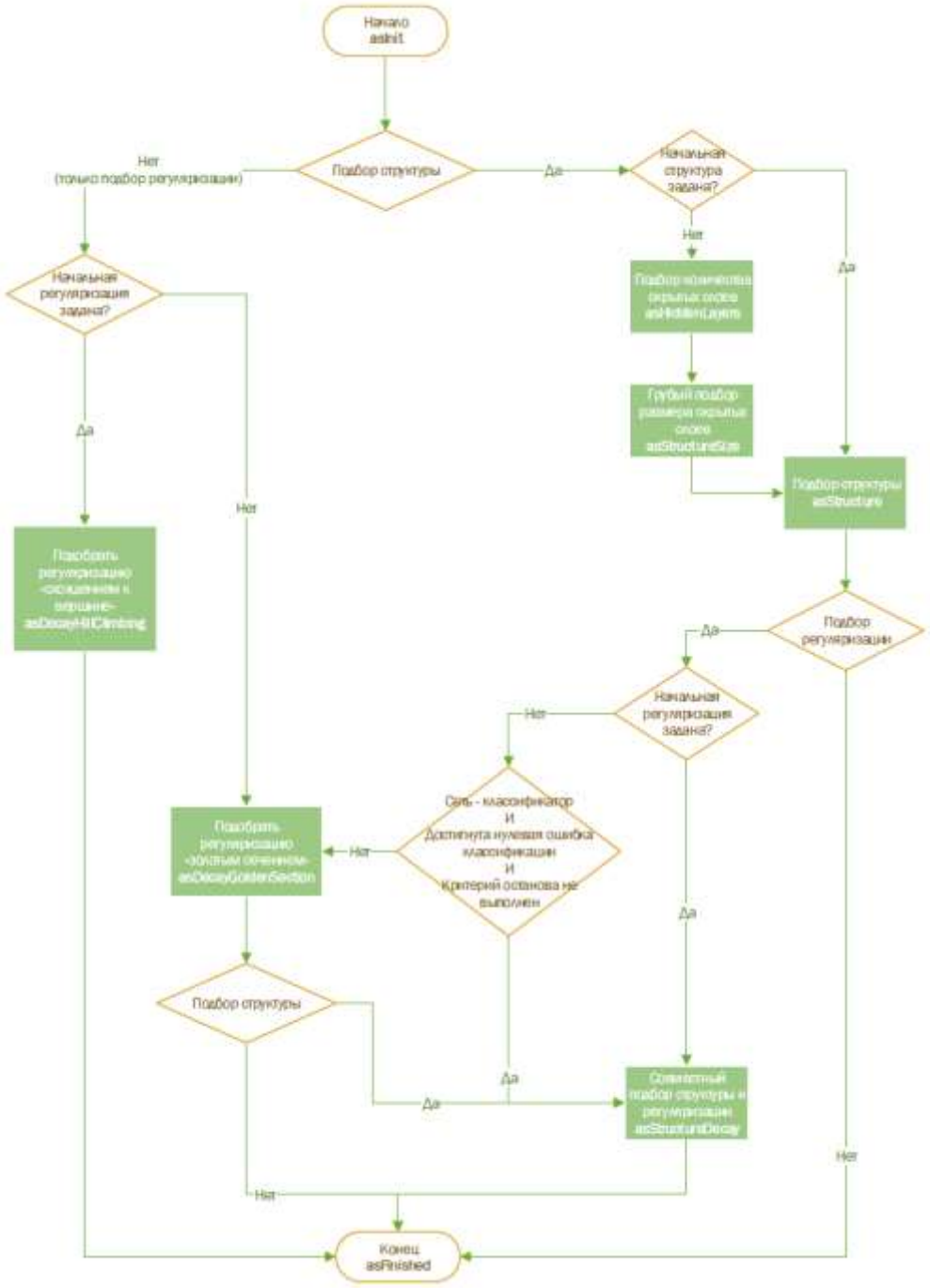
Конечные классы:

1. Все показатели на одном экране
2. Ручная правка разбиения
3. Поддержка внешних диапазонов

Нейросети

Автоподбор структуры и степени регуляризации:

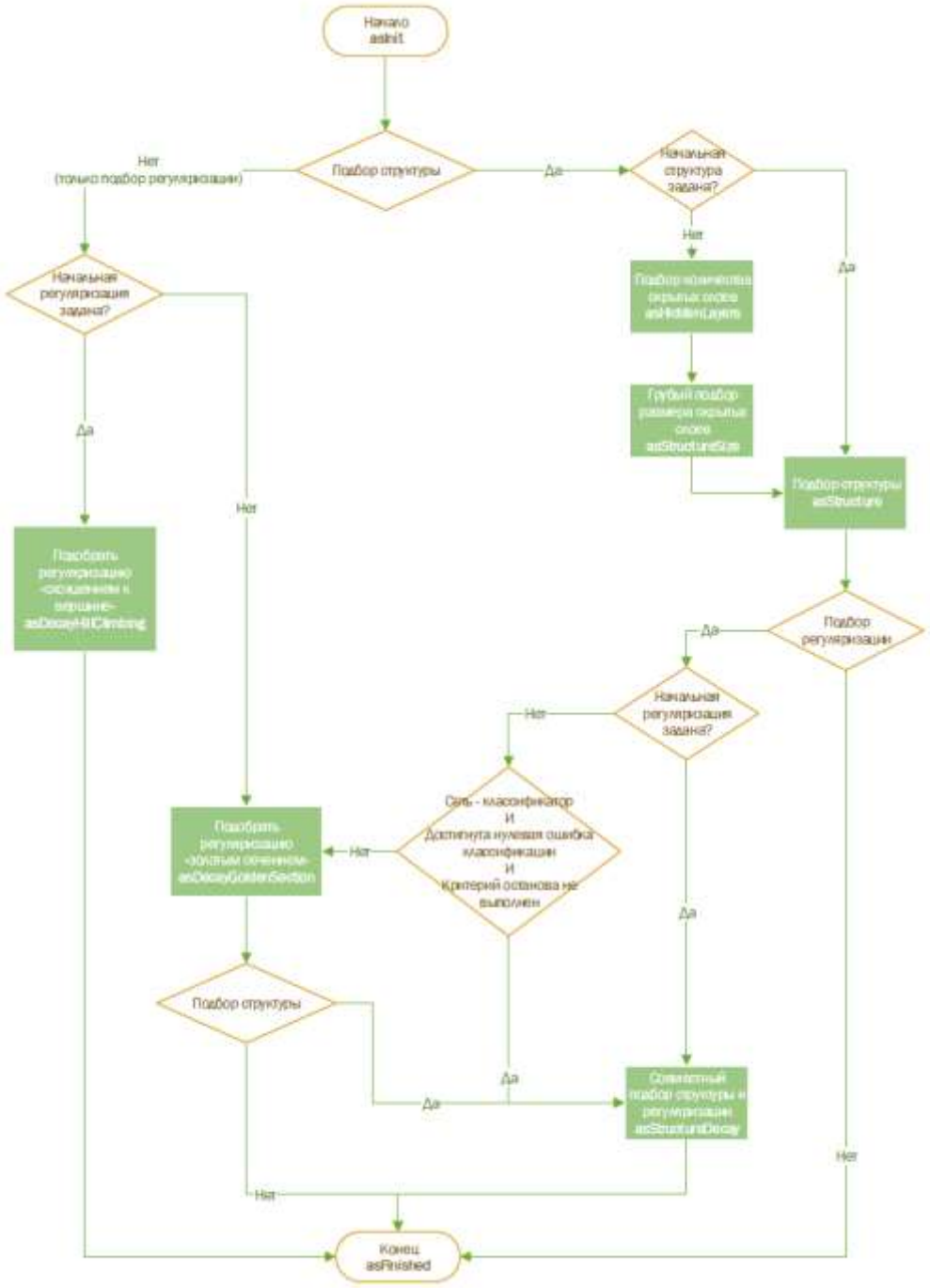
- Оптимальная структура
- Защита от переобучения
- Экономия времени и памяти



Нейросети

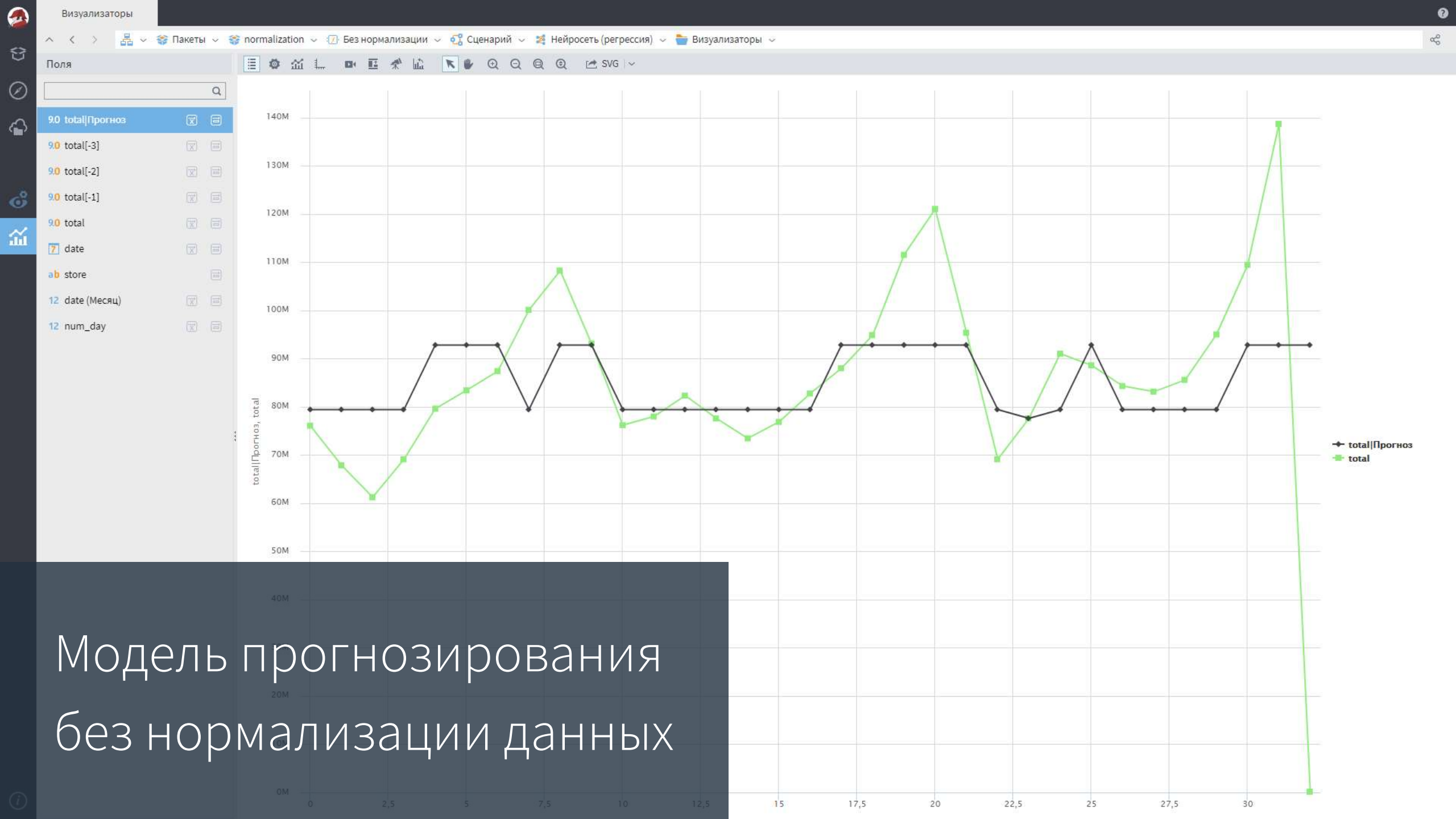
Метаалгоритм, не требующий участия пользователя с начальной точкой и без:

- Golden Section
- Hill Climbing
- Exhaustive Search



Отсутствие нормализации (денормализации) усложняет работу аналитика:

- Плохо обусловленные данные
- Долгое схождение алгоритмов
- Неадекватные результаты



Модель прогнозирования без нормализации данных

Сложности ручной нормализации

- Самостоятельный расчет, сохранение, поддержание и синхронизация настроек
- Поддержание списков уникальных значений
- Трудности при замене входных данных
- Цена ошибки – неработающая модель

Нормализация/денормализация

Непрерывные данные

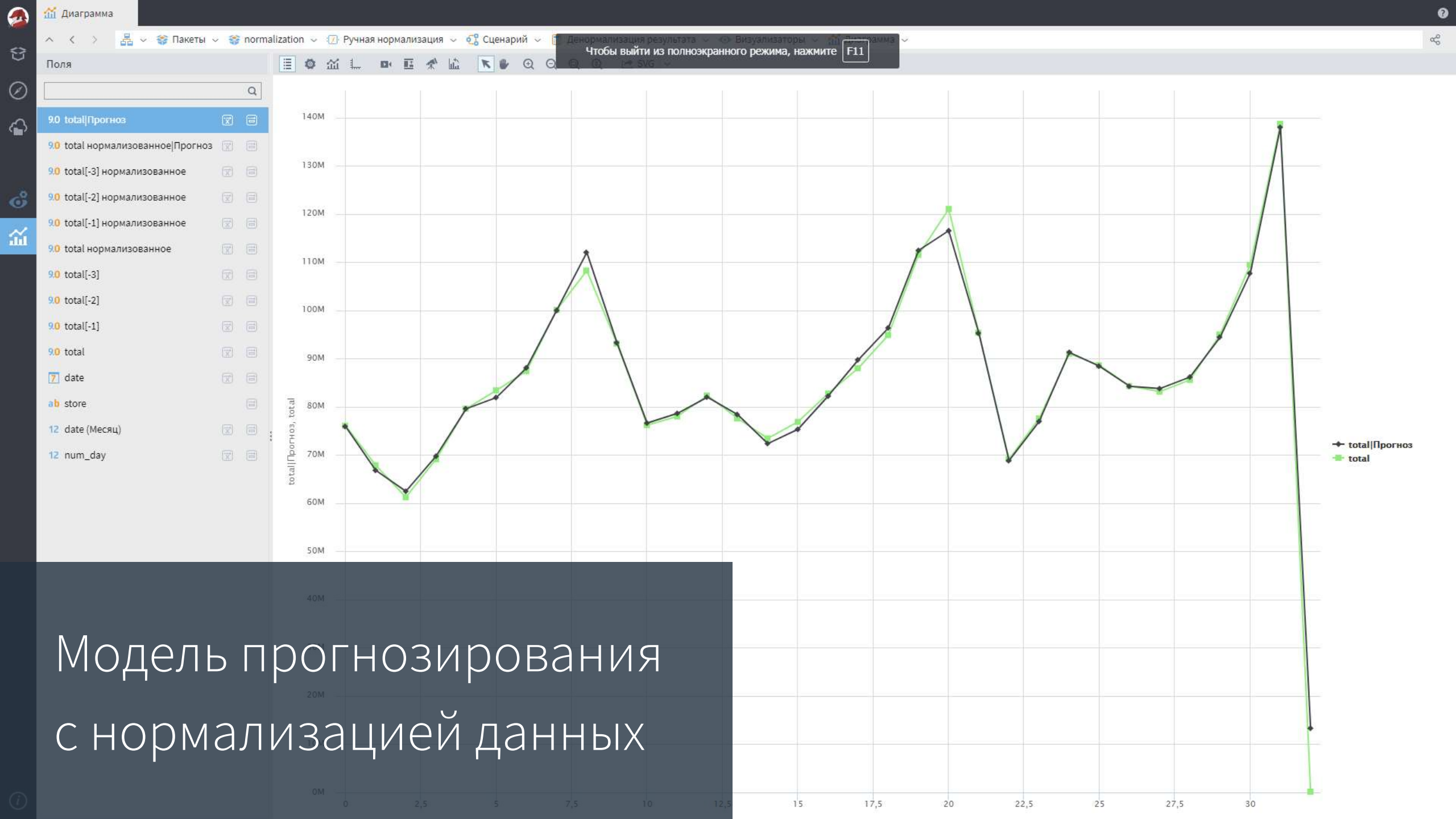
1. Нормализация MIN-MAX
2. Нормализация [0; 1]
3. Нормализация [-1; 1]
4. Абсолютное масштабирование
5. Стандартизация
6. Отношение

Дискретные данные

1. Индикатор
2. Индикатор без опорной категории
3. Отклонение
4. Простой
5. Разность
6. Обратная разность
7. Helmert
8. Обратный Helmert
9. Индекс уникального значения

Нормализация/денормализация встроена в Data Mining алгоритмы:

- Логистическая регрессия
- Линейная регрессия
- Нейросеть (классификация, регрессия)
- Кластеризация (k-means)
- EM-кластеризация
- SONN
- Прогнозирование (ARIMAX)



Модель прогнозирования с нормализацией данных

Обобщающая способность моделей

Разбиение на множества встроено в компоненты там, где оно необходимо:

- Случайный и последовательный сэмплинг
- Автоматический расчет всех показателей на обучающем и тестовом множествах

Обобщающая способность моделей

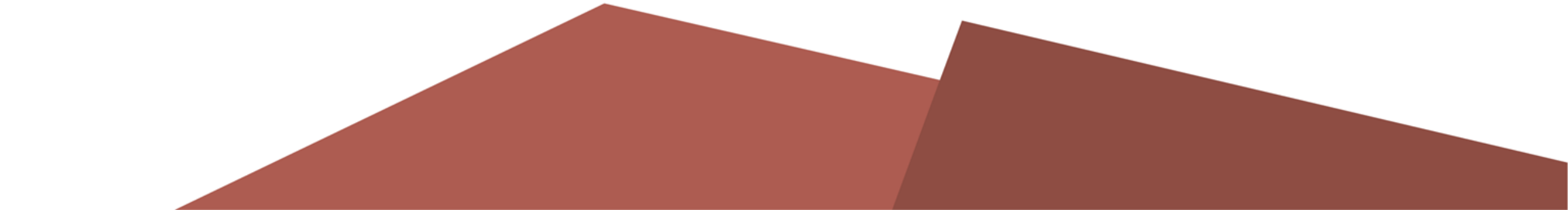
Кросс-валидация встроена в компоненты, допускающие подобную проверку:

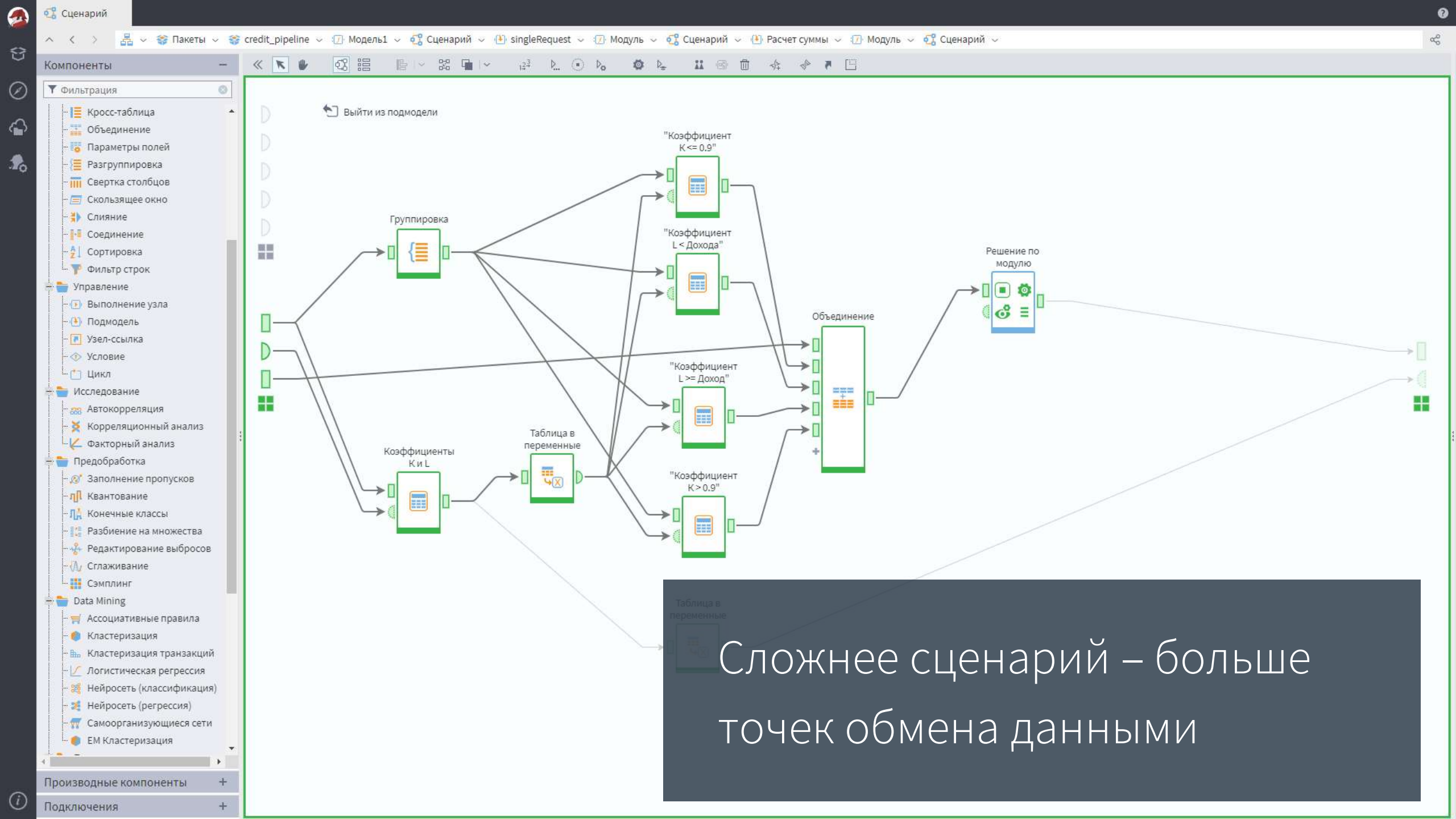
- K-folds
- Shuffle split

Обобщающая способность моделей

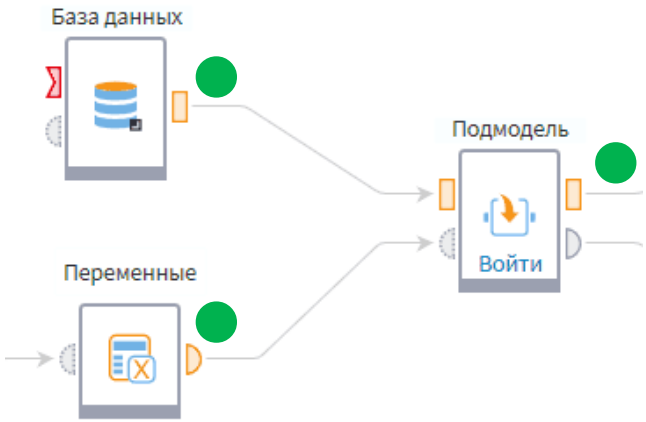
- Тестирование и кросс-валидация без дополнительных усилий
- Выходные показатели унифицированы среди различных компонентов
- Удобное сравнение моделей различной природы

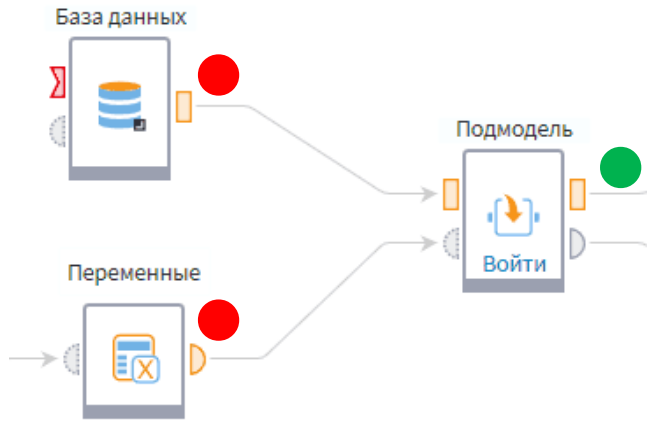
Удобство проектирования

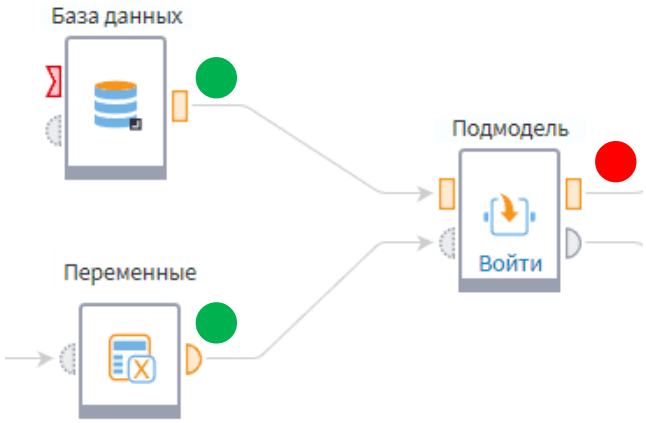


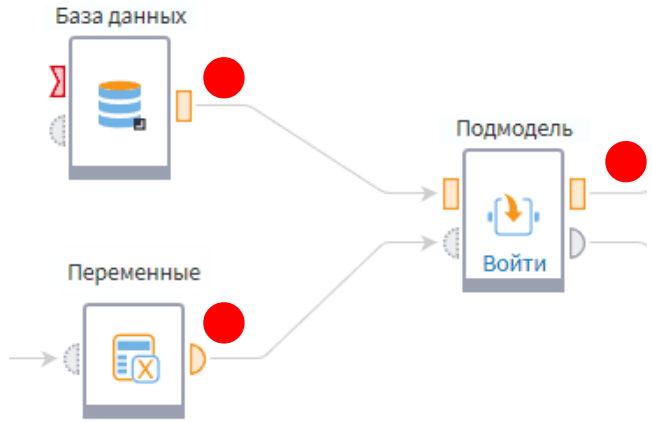

Сценарий
Пакеты
credit_pipeline
Модель1
Сценарий
singleRequest
Модуль
Сценарий
Расчет суммы
Модуль
Сценарий
Компоненты
Фильтрация
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Управление
Выполнение узла
Подмодель
Узел-ссылка
Условие
Цикл
Исследование
Автокорреляция
Корреляционный анализ
Факторный анализ
Предобработка
Заполнение пропусков
Квантование
Конечные классы
Разбиение на множества
Редактирование выбросов
Сглаживание
Сэмплинг
Data Mining
Ассоциативные правила
Кластеризация
Кластеризация транзакций
Логистическая регрессия
Нейросеть (классификация)
Нейросеть (регрессия)
Самоорганизующиеся сети
EM Кластеризация
Производные компоненты
Подключения
Выйти из подмодели
"Коэффициент K <= 0.9"
Группировка
"Коэффициент L < Дохода"
Решение по модулю
Объединение
"Коэффициент L >= Доход"
Таблица в переменные
Коэффициенты K и L
"Коэффициент K > 0.9"
Сложнее сценарий – больше точек обмена данными

Автосинхронизация: кейсы









Автосинхронизация

Обязательные поля:

- Настроенные пользователем
- С заданным назначением

Необязательные поля:

- Пробрасываем насквозь
- Данные – только по требованию

Автосинхронизация

- Минимум действий пользователя при подмене источников данных
- Все конфликты подсвечиваем визуально
- Покажем мастер, только если потребуется вмешательство пользователя

Эволюция Loginom

1. Больше алгоритмов анализа
2. Повышение качества анализа
3. Автоматизация рутинной работы
4. Улучшение интерпретируемости

loginom.ru