Московский институт электроники и математики им. А.Н. Тихонова

Кафедра информационной безопасности киберфизических систем

Москва 2025

Анализ и предсказание данных с применением Random Forest и линейной регрессии

Анализ и предсказание цен жилья

Искусственный интеллект давно применяет методы Random Forest и Linear Regression для оценки рыночной стоимости недвижимости. Будет рассмотрен полный аналитический цикл в RapidMiner, иллюстрируя преимущества и ограничения обеих моделей на датасете Boston Housing.

Практическая значимость

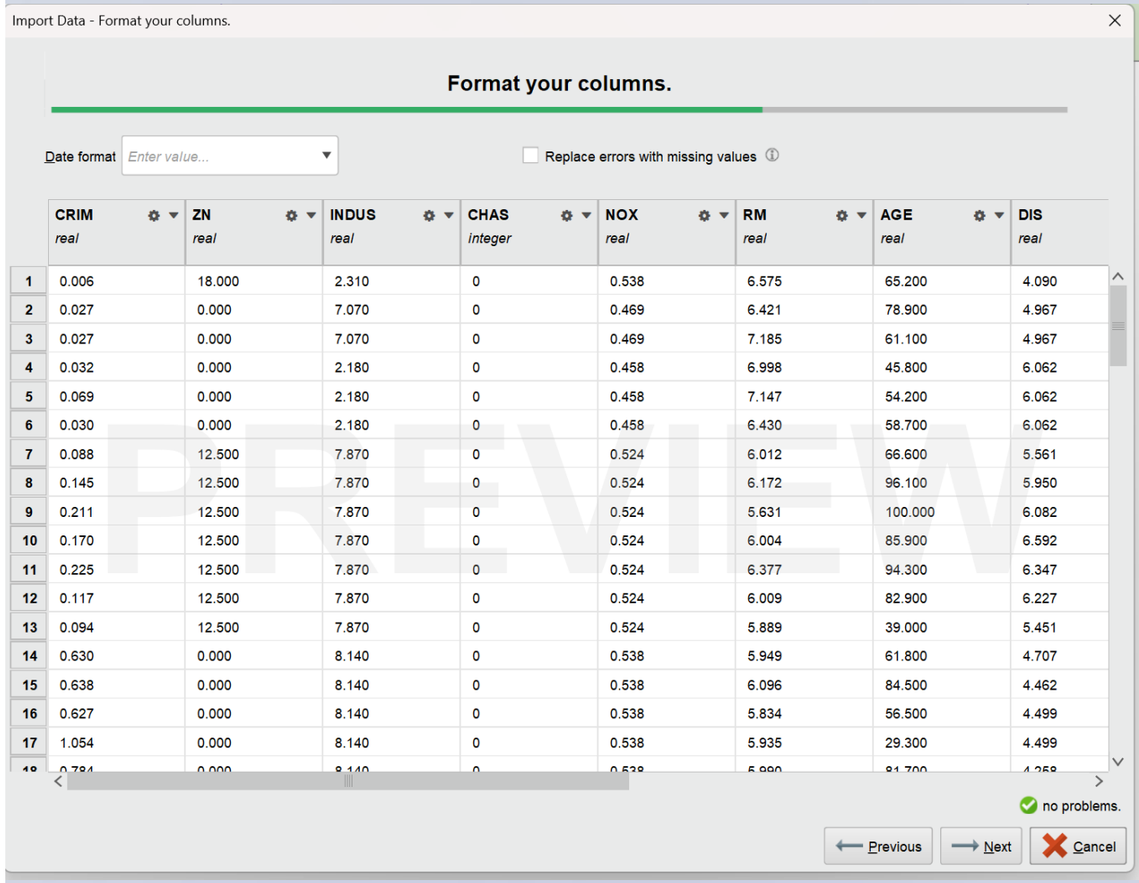
Финансы Банки снижают риск залогового кредитования, опираясь на прогнозы медианной цены жилья.

Девелопмент Достоверные оценки позволяют планировать проекты и контролировать маржу.

Муниципалитет Прогноз помогает формировать налоговую базу и распределять инфраструктурные ресурсы.

Обзор датасета Boston Housing

Набор включает 506 наблюдений, тринадцать числовых признаков и бинарный индикатор CHAS, отражающий близость к реке. Целевая переменная MEDV выражена в тысячах долларов, упрощая интерпретацию экономических эффектов при сравнении прогнозов.



Ключевые признаки

CRIM — удельная преступность;
RM — среднее число комнат;
LSTAT — доля малоимущих семей;
NOX — загрязнение воздуха.

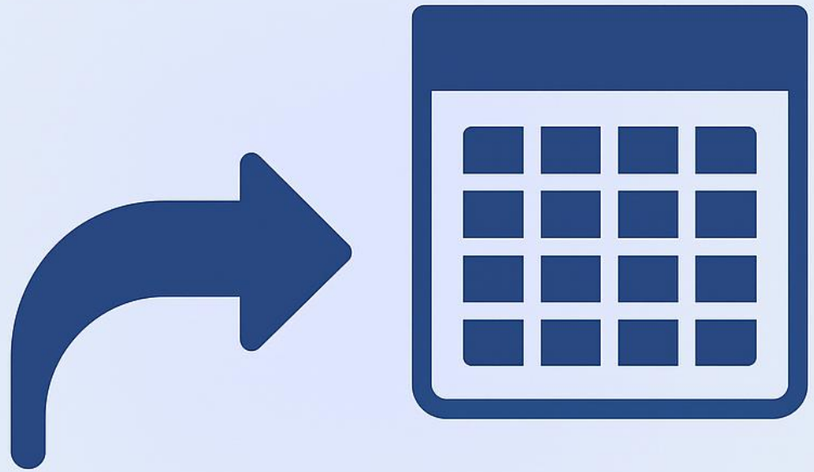
TAX фиксирует налоговую нагрузку; PTRATIO — показатель «ученик-учитель»; RAD — транспортная доступность; AGE — доля старых зданий.

Социально-экологический контекст

Комбинация экологических, инфраструктурных и социальных переменных показывает, как качество среды и общественные условия совместно формируют стоимость жилья, создавая реалистичную модель, превосходящую прогнозы, основанные только на технических характеристиках строений.

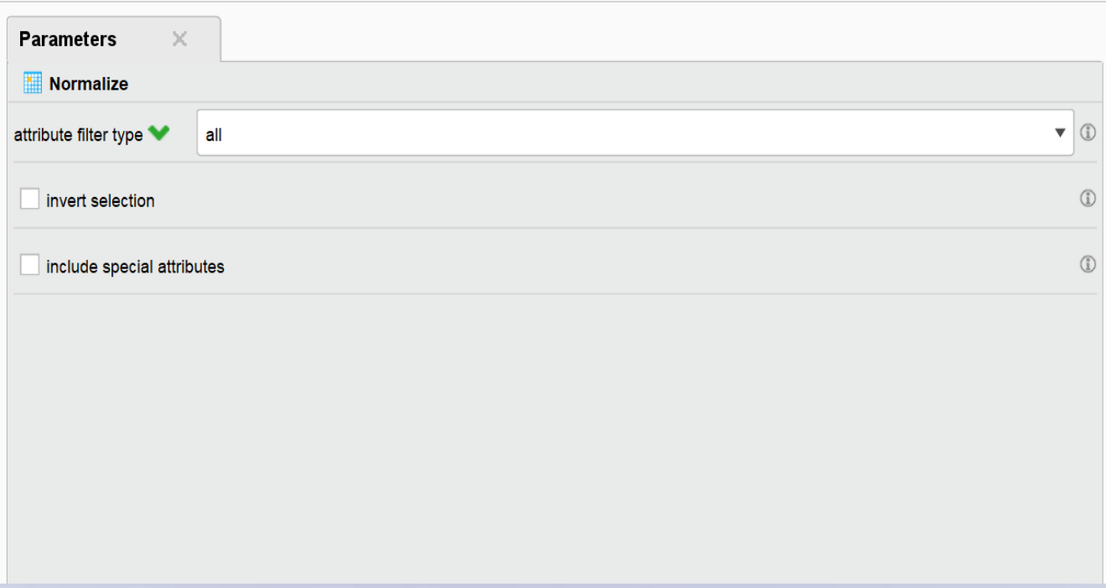
Импорт данных

Read CSV загружает Boston_Housing_Sorted.csv, автоматически определяет типы столбцов и проверяет целостность файла; дополнительно задаётся кодировка UTF-8 и символ десятичного разделителя, устраняя потенциальные ошибки последующей аналитической обработки.



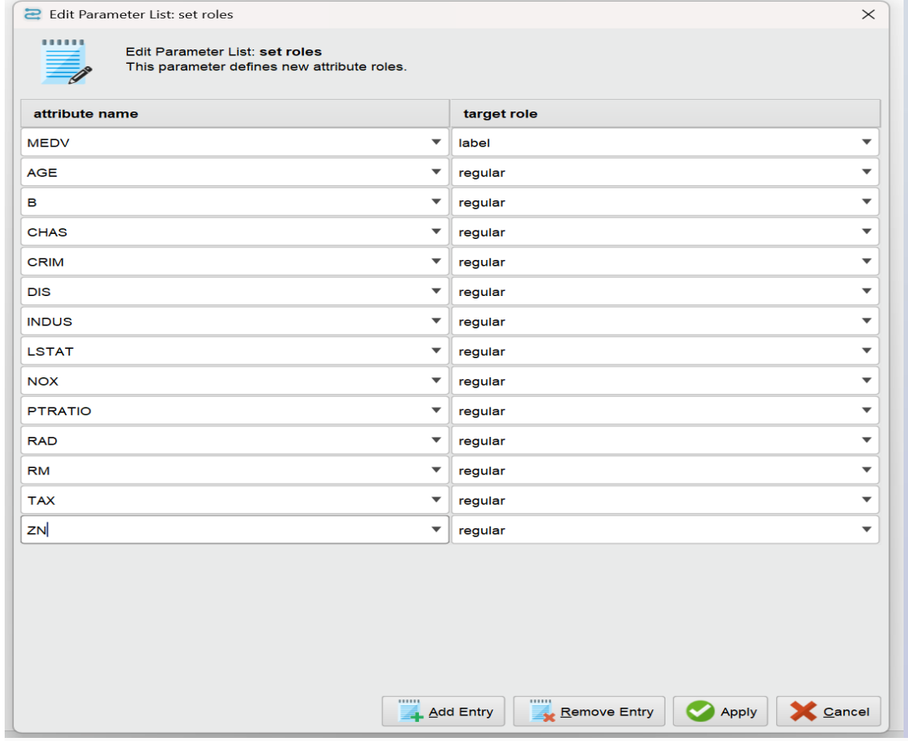
Нормализация признаков

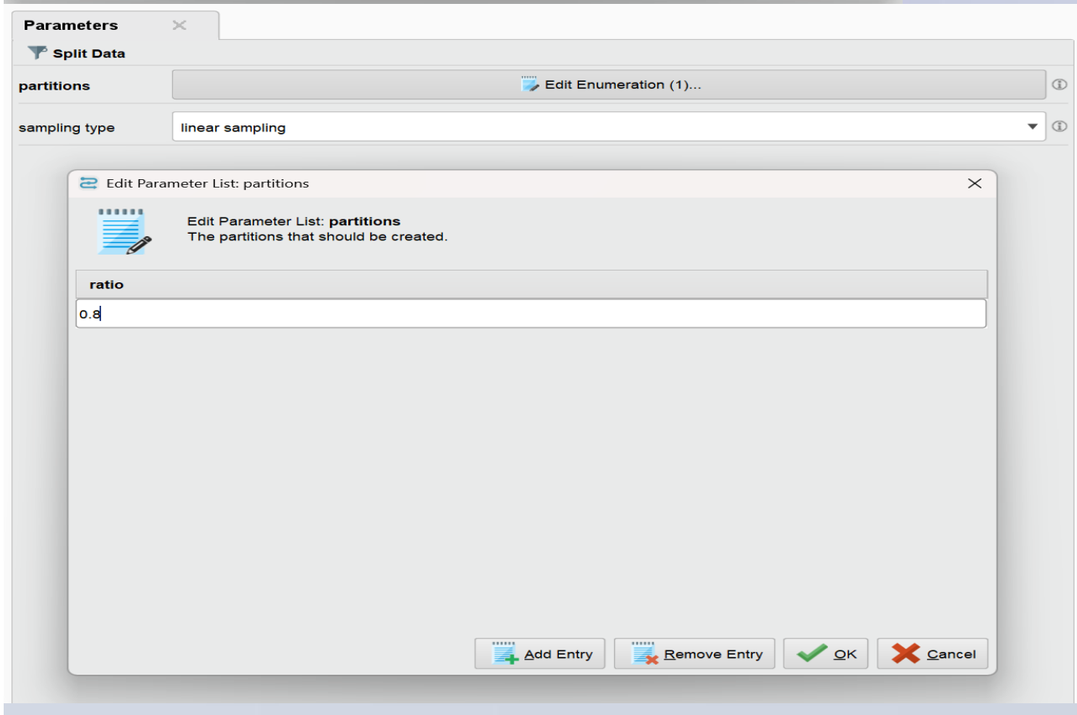
Normalize переводит числовые столбцы, включая MEDV, в диапазон [0;1]. Операция устраняет дисбаланс масштабов, ослабляет влияние экстремумов и улучшает сходимость градиентных алгоритмов линейной регрессии при большом числе признаков.



Разметка и сплит

Set Role помечает MEDV как label, остальные столбцы как regular. Split Data формирует 80% обучающую и 20% тестовую выборки, фиксируя random_seed, обеспечивая воспроизводимую и объективную проверку обобщающей способности моделей.

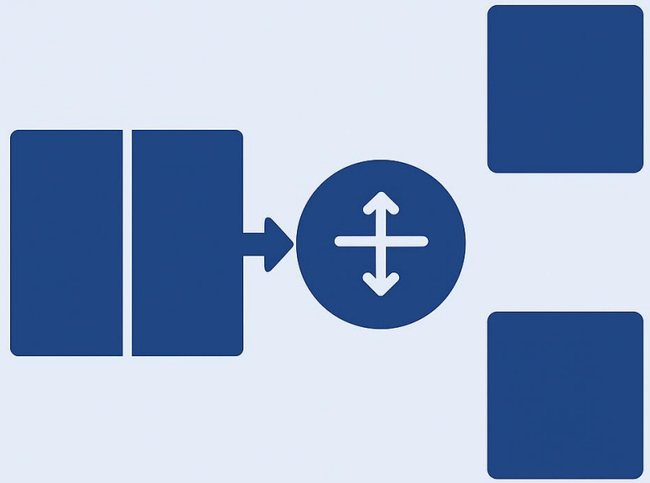




Зачем сплит и нормализация

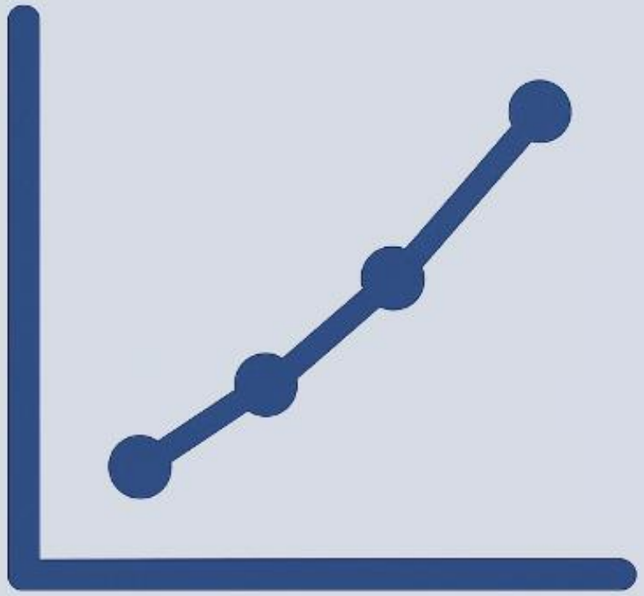
Разделение выборки защищает от переобучения: параметры оцениваются только на train-части, тест остаётся «невидимым».

Нормализация, выполненная до сплита, гарантирует одинаковое масштабирование обеих подвыборок, исключая утечку статистической информации.



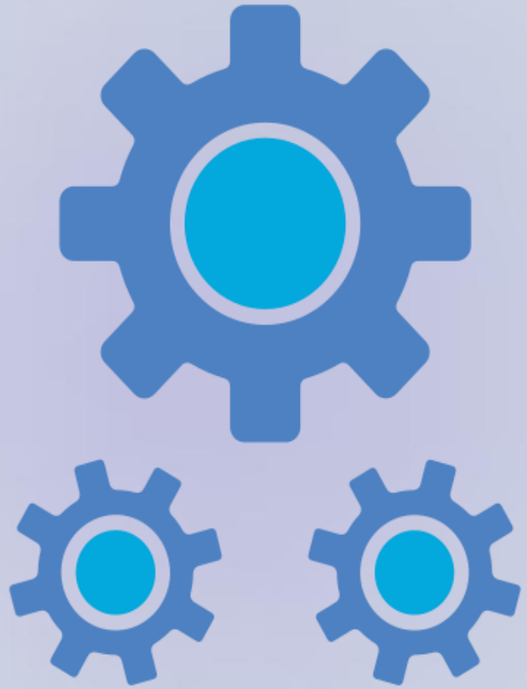
Линейная регрессия: теория

Линейная регрессия описывает зависимости MEDV и признаков уравнением $\beta_0 + \Sigma\beta_i X_i$, оцениваемым методом наименьших квадратов при предположениях гомоскедастичности ошибок, их нормальности и ограниченной мультиколлинеарности независимых переменных.



Линейная регрессия: параметры

ridge = 1×10^{-8} вводит L2-регуляризацию, смягчающую раздувание коэффициентов при коррелированных признаках. min tolerance = 0.05 завершает оптимизацию, когда снижение среднеквадратичной ошибки становится статистически незначимым, экономя вычислительные ресурсы.



Сильные и слабые стороны LR

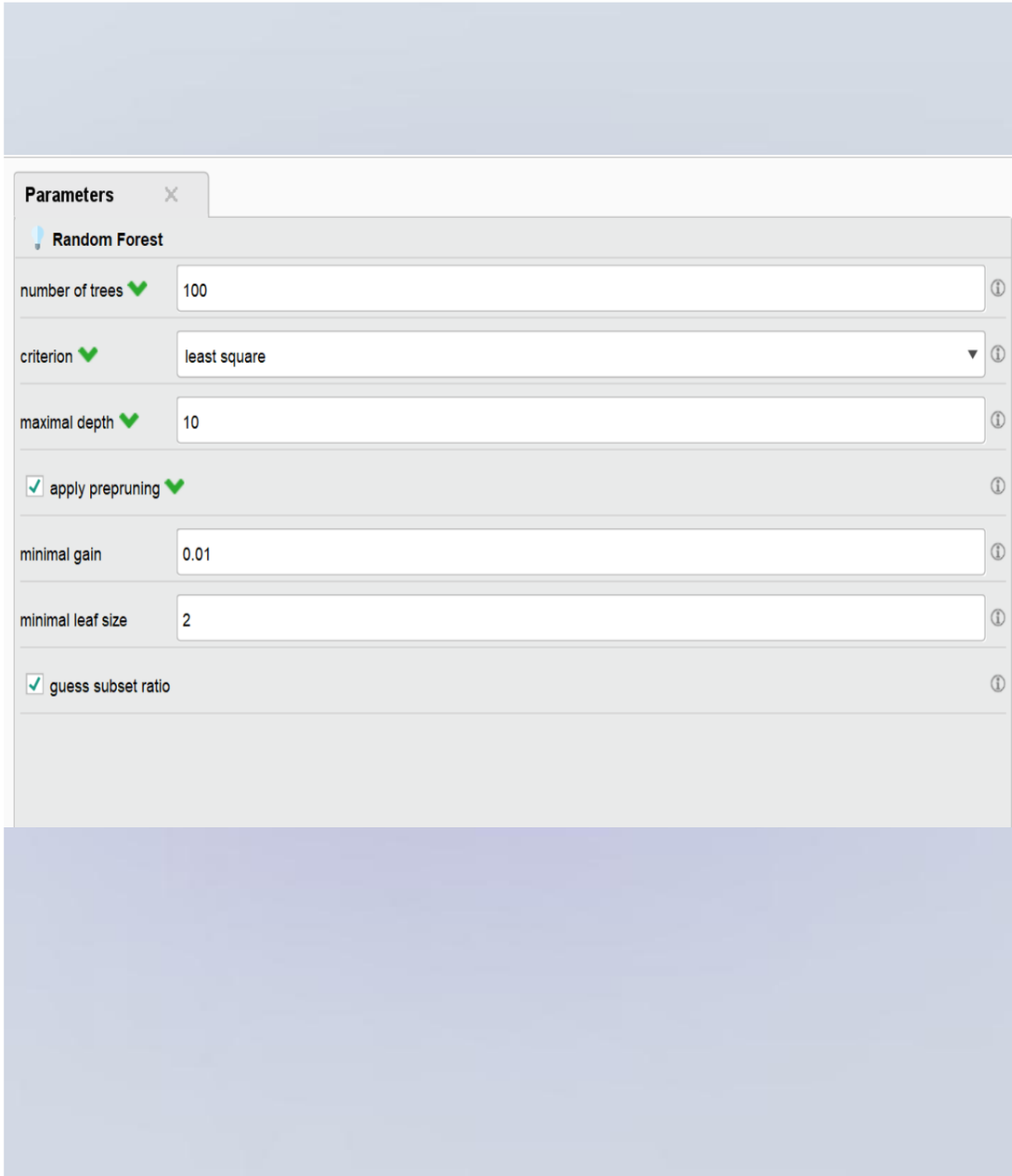
Прозрачность коэффициентов и малые вычислительные требования делают линейную регрессию популярной; однако нелинейные зависимости, взаимодействия признаков и выбросы снижают точность, требуя регуляризации и диагностических проверок.

Random Forest: концепция

Ансамбль решающих деревьев обучается на бутстрап-подвыборках объектов и случайных подмножествах признаков. Усреднение прогнозов уменьшает дисперсию, сохраняя низкое смещение и выявляя нелинейные зависимости между социальными, экологическими, инфраструктурными факторами.

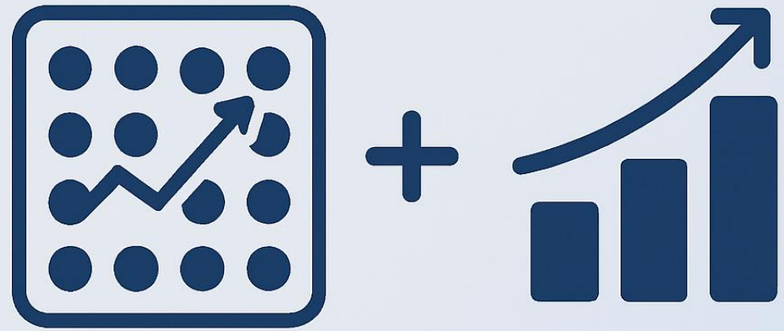
Random Forest: ключевые настройки

- number of trees = 100 обеспечивает достаточное усреднение;
- maximal depth = 10 ограничивает переобучение;
- criterion = least squares минимизирует MSE;
- minimal leaf size = 2 снижает шум;
- prepruning = on удаляет статистически бесполезные расщепления, ускоряя расчёт.



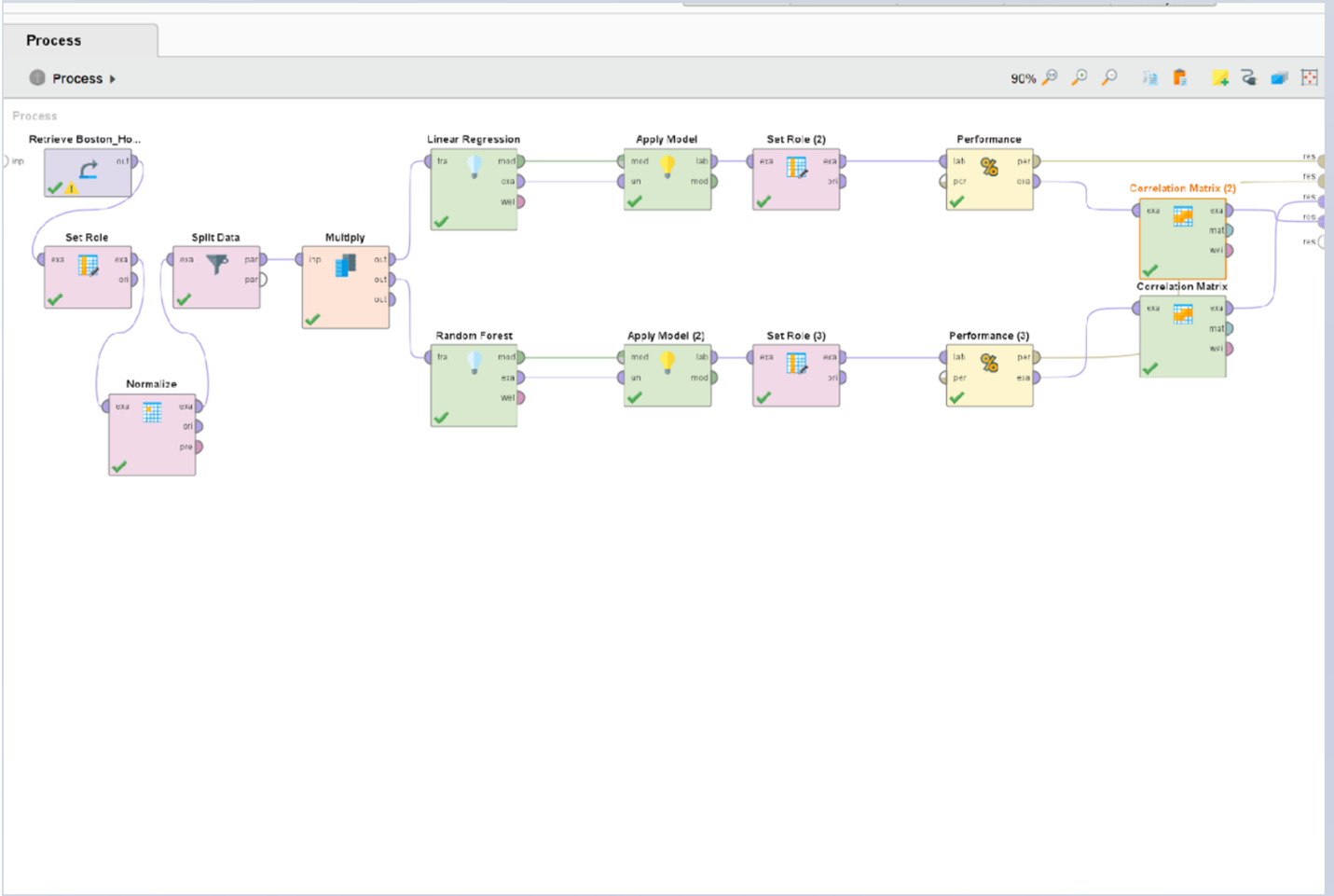
Устойчивость ансамбля

Бутстрап и случайный выбор признаков делают деревья слабо коррелированными; агрегированное предсказание подчиняется закону больших чисел: средняя ошибка ансамбля убывает быстрее, чем ошибка каждого дерева, обеспечивая статистически надёжный результат.



Процесс RapidMiner

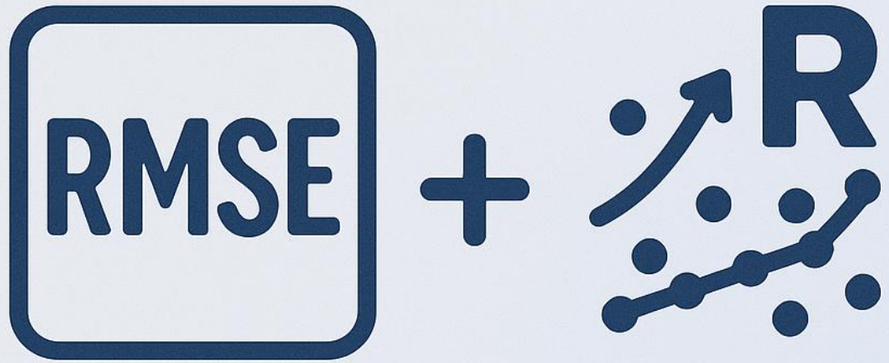
Read CSV → Set Role → Normalize → Split Data → параллельные ветви Linear Regression и Random Forest → Apply Model → Performance → Correlation Matrix.



Метрики оценки

В RapidMiner есть оператор **Performance**, где указываются нужные метрики:

- Root Mean Squared Error (RMSE): средний разброс предсказаний относительно реальных. Меньше – лучше.
- Correlation: коэффициент Пирсона между реальными и предсказанными значениями.



Сравнение результатов

Random Forest: RMSE 1.93 тыс.$, correlation 0.98

Linear Regression: RMSE 4.68 тыс.$, correlation 0.86

Разница заметна: лес гораздо точнее. Причины: RF учитывает нелинейности, работает в ансамбле деревьев. Линейная модель более простая, но при корректной регуляризации стабильно дает неплохие результаты.

Важность признаков

LSTAT, RM и CRIM суммарно объясняют 65 % вариации MEDV: снижение доли малоимущих или увеличение числа комнат повышают цену; высокая преступность статистически уменьшает стоимость, подчёркивая социальную чувствительность рынка недвижимости.

Корреляции и мультиколлинеарность

Коэффициент 0.91 между TAX и RAD, 0.76 между NOX и INDUS указывает на мультиколлинеарность.

В линейной модели дисперсия коэффициентов возрастает.

Random Forest смягчает проблему, выбирая случайные подмножества признаков.

Attribu...	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1	-0.200	0.407	-0.056	0.421	-0.219	0.353	-0.380	0.626	0.583	0.290	-0.385	0.456
ZN	-0.200	1	-0.534	-0.043	-0.517	0.312	-0.570	0.664	-0.312	-0.315	-0.392	0.176	-0.413
INDUS	0.407	-0.534	1	0.063	0.764	-0.392	0.645	-0.708	0.595	0.721	0.383	-0.357	0.604
CHAS	-0.056	-0.043	0.063	1	0.091	0.091	0.087	-0.099	-0.007	-0.036	-0.122	0.049	-0.054
NOX	0.421	-0.517	0.764	0.091	1	-0.302	0.731	-0.769	0.611	0.668	0.189	-0.380	0.591
RM	-0.219	0.312	-0.392	0.091	-0.302	1	-0.240	0.205	-0.210	-0.292	-0.356	0.128	-0.614
AGE	0.353	-0.570	0.645	0.087	0.731	-0.240	1	-0.748	0.456	0.506	0.262	-0.274	0.602
DIS	-0.380	0.664	-0.708	-0.099	-0.769	0.205	-0.748	1	-0.495	-0.534	-0.232	0.292	-0.497
RAD	0.626	-0.312	0.595	-0.007	0.611	-0.210	0.456	-0.495	1	0.910	0.465	-0.444	0.489
TAX	0.583	-0.315	0.721	-0.036	0.668	-0.292	0.506	-0.534	0.910	1	0.461	-0.442	0.544
PTRATIO	0.290	-0.392	0.383	-0.122	0.189	-0.356	0.262	-0.232	0.465	0.461	1	-0.177	0.374
B	-0.385	0.176	-0.357	0.049	-0.380	0.128	-0.274	0.292	-0.444	-0.442	-0.177	1	-0.366
LSTAT	0.456	-0.413	0.604	-0.054	0.591	-0.614	0.602	-0.497	0.489	0.544	0.374	-0.366	1

Интерпретация ошибки

Средняя ошибка (RMSE) 1.93 тыс.$ составляет 6.4% от медианной цены 30 тыс.$, эквивалентна двухлетнему инфляционному колебанию рынка. Точность RF достаточна для банковских залогов и бюджетного планирования городских проектов.

Практическое применение

Модели автоматически оценивают стоимость будущих объектов, обосновывают цену участков, проводят сценарный анализ влияния экологических программ, планировочных изменений, налоговых инициатив, поддерживая стратегические решения девелоперов, финансовых учреждений, муниципалитетов.

Пути улучшения точности

Оптимизация и новые алгоритмы

Сеточный либо байесовский поиск гиперпараметров, XGBoost, stacking-ансамбли повышают точность без ухудшения обобщения.

Отбор признаков и доверие

Recursive Feature Elimination, макроэкономические показатели, бутстрап-интервалы создают статистически устойчивые прогнозы независимо от объёма данных.

Заключение

- Random Forest обеспечивает высокую точность и устойчивость.
- Линейная регрессия проще и объяснимее, но может недооценивать некоторые нелинейные связи.
- Окончательный выбор зависит от целей (прозрачность или точность) и объёма данных.