Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

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

ПРАКТИЧЕСКАЯ РАБОТА № 11
по дисциплине «Информатика»

ТЕМА РАБОТЫ

«Анализ и предсказание данных с применением Random Forest и линейной регрессии»

Москва, 2024

Оглавление

1. Введение

Целью данной лабораторной работы является освоение методов анализа данных и машинного обучения с использованием алгоритмов Random Forest и Linear Regression с помощью инструмента RapidMiner. Эти методы машинного обучения позволяют строить модели для прогнозирования значений целевой переменной на основе различных признаков.

В рамках работы студенты выполнят предобработку данных, обучат модели Random Forest и Linear Regression, проведут анализ важности признаков и оценят качество моделей с использованием метрик Root Mean Squared Error (RMSE) и Correlation.

Кроме того, в ходе работы будет проведен анализ влияния различных признаков на целевую переменную, а также визуализация результатов предсказаний с использованием графиков для более наглядной интерпретации работы моделей.

2. Содержание практической работы

Описание работы:

Анализ проводится на наборе данных о недвижимости, включающем атрибуты, такие как количество комнат, уровень преступности, стоимость недвижимости, возраст здания и другие факторы. Основная цель работы — построение модели для предсказания цены недвижимости на основе этих факторов с использованием алгоритмов Random Forest и Linear Regression.

Этапы выполнения работы:

1. Провести предобработку данных, включая удаление ненужных столбцов и выбор значимых атрибутов для анализа.
2. Разделить данные на обучающую и тестовую выборки с использованием метода "Split Data".
3. Построить модель Random Forest, настроить параметры.
4. Построить модель Linear Regression, настроить параметры.
5. Применить обученные модели для предсказания стоимости недвижимости.
6. Оценить качество моделей, использую RMSE и correlation.
7. Провести анализ важности признаков, влияющих на итоговую оценку, и визуализировать их значимость.
8. Построить корреляционные матрицы.

О наборе данных:

Анализ проводится на наборе данных, содержащем следующие характеристики:

- **CRIM** (уровень преступности на душу населения),
- **ZN** (доля земли, предназначенной для жилой застройки),
- **INDUS** (доля индустриальных площадей),
- **CHAS** (наличие реки поблизости),
- **NOX** (концентрация оксидов азота в воздухе),
- **RM** (среднее количество комнат в доме),

- **AGE** (доля зданий, построенных до 1940 года),
- **DIS** (дистанция до центров занятости),
- **RAD** (доступность транспорта),
- **TAX** (налоговая ставка на недвижимость),
- **PTRATIO** (соотношение учеников и учителей по городу),
- **B** (доля чернокожих жителей в районе),
- **LSTAT** (процент людей с низким социальным статусом),
- **MEDV** (медианная стоимость жилья, выраженная в тысячах долларов — целевая переменная).

Ключевые особенности данных:

Количество записей: содержит данные о 506 объектах недвижимости.

Формат данных: табличный набор, включающий числовые и категориальные переменные.

Потенциальные проблемы:

Различие форматов признаков (числовые и категориальные данные). Невозможность учета всех социальных и экономических факторов, которые могут влиять на цену недвижимости.

Специфика работы:

Для выполнения лабораторной работы используются будет использован весь датасет. Эти данные позволят проанализировать, какие факторы наиболее сильно влияют на цену недвижимости и построить предсказательную модель для оценки стоимости жилья.

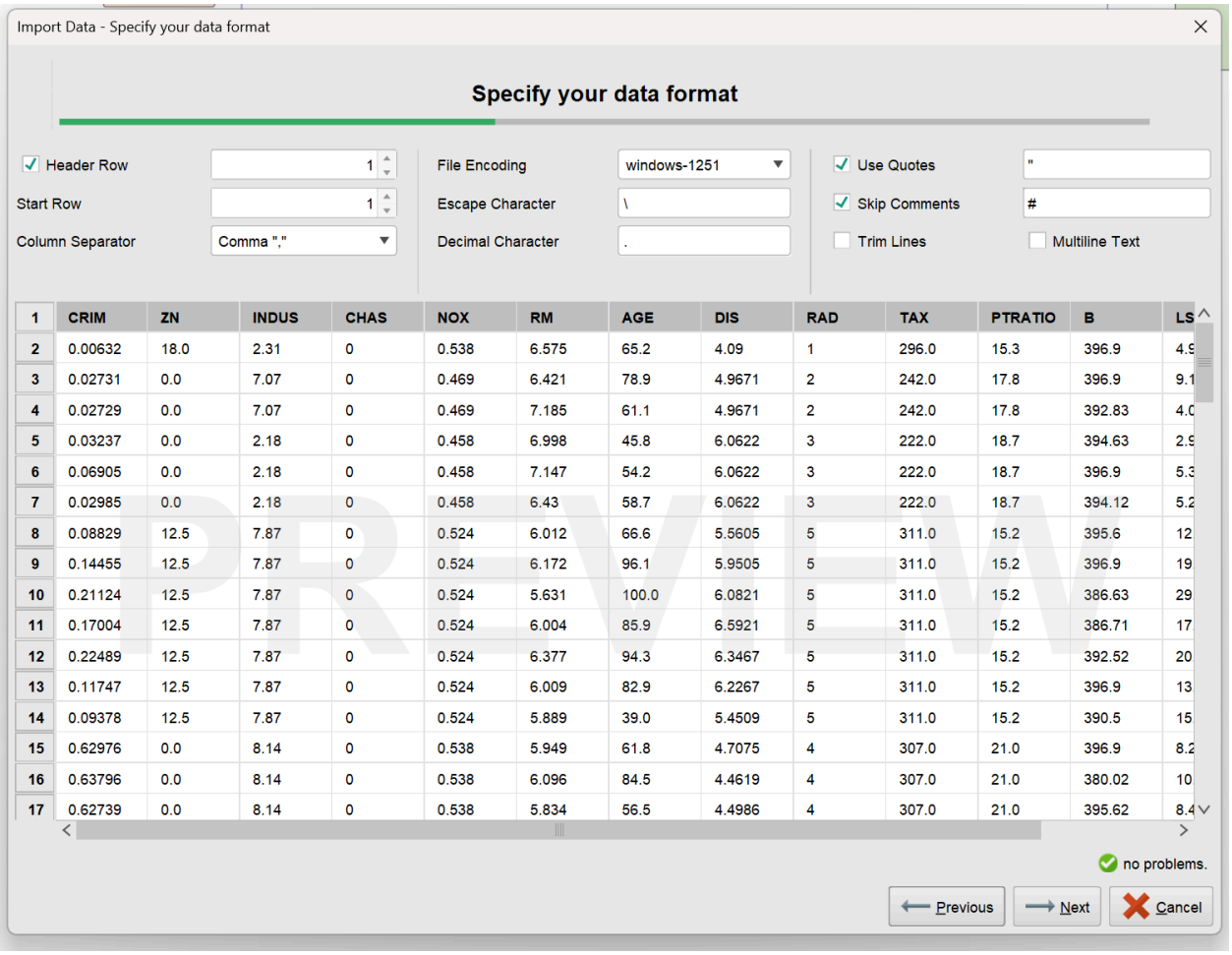
3. Ход работы

Загрузка набора данных

1. Откройте RapidMiner Studio.
2. В главном меню выберите **"Create New Process"**.
3. Воспользуйтесь функцией **"Import data".**

4. Загрузите набор данных о транзакциях, выбрав файл **"Boston_Housing_Sorted.csv"** в формате csv.
5. Сохраните полученную базу данных в папку со своей работой.
6. В результате вы увидите таблицу с данными, содержащими атрибуты: (**CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV**)

Данные успешно загружены, их структура показана на рисунке 3.2.



1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LS
2	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.09	1	296.0	15.3	396.9	4.9
3	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.9	9.1
4	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.0
5	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.9
6	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.9	5.3
7	0.02985	0.0	2.18	0	0.458	6.43	58.7	6.0622	3	222.0	18.7	394.12	5.2
8	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311.0	15.2	395.6	12
9	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311.0	15.2	396.9	19
10	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311.0	15.2	386.63	29
11	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311.0	15.2	386.71	17
12	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311.0	15.2	392.52	20
13	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311.0	15.2	396.9	13
14	0.09378	12.5	7.87	0	0.524	5.889	39.0	5.4509	5	311.0	15.2	390.5	15
15	0.62976	0.0	8.14	0	0.538	5.949	61.8	4.7075	4	307.0	21.0	396.9	8.2
16	0.63796	0.0	8.14	0	0.538	6.096	84.5	4.4619	4	307.0	21.0	380.02	10
17	0.62739	0.0	8.14	0	0.538	5.834	56.5	4.4986	4	307.0	21.0	395.62	8.4

Рисунок 3.1 – подготовка данных к выгрузке

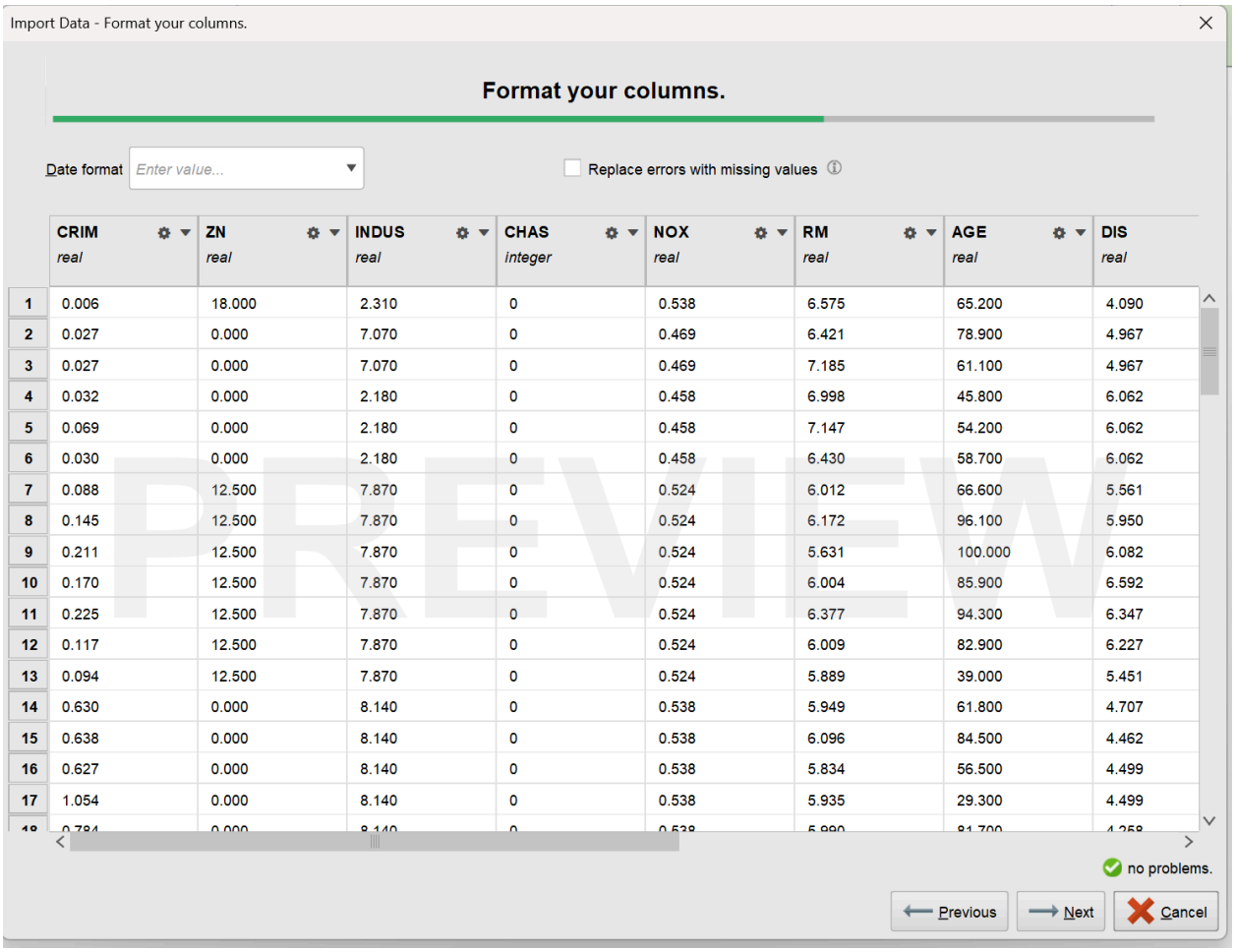


Рисунок 3.2 – выгруженные данные

Нормализация данных:

На этом этапе мы будем нормализовать данные с помощью блока **"Normalize"**. Нормализация помогает привести все признаки к единому масштабу, что важно для некоторых моделей машинного обучения, чтобы избежать доминирования признаков с большими значениями.

В блоке выберите "attribute filter type" как "all", чтобы применить нормализацию ко всем признакам, включая числовые значения и целевую переменную MEDV. Это гарантирует, что каждый признак будет иметь одинаковый масштаб, что улучшит работу моделей. Финальные настройки представлены на рисунке 3.3

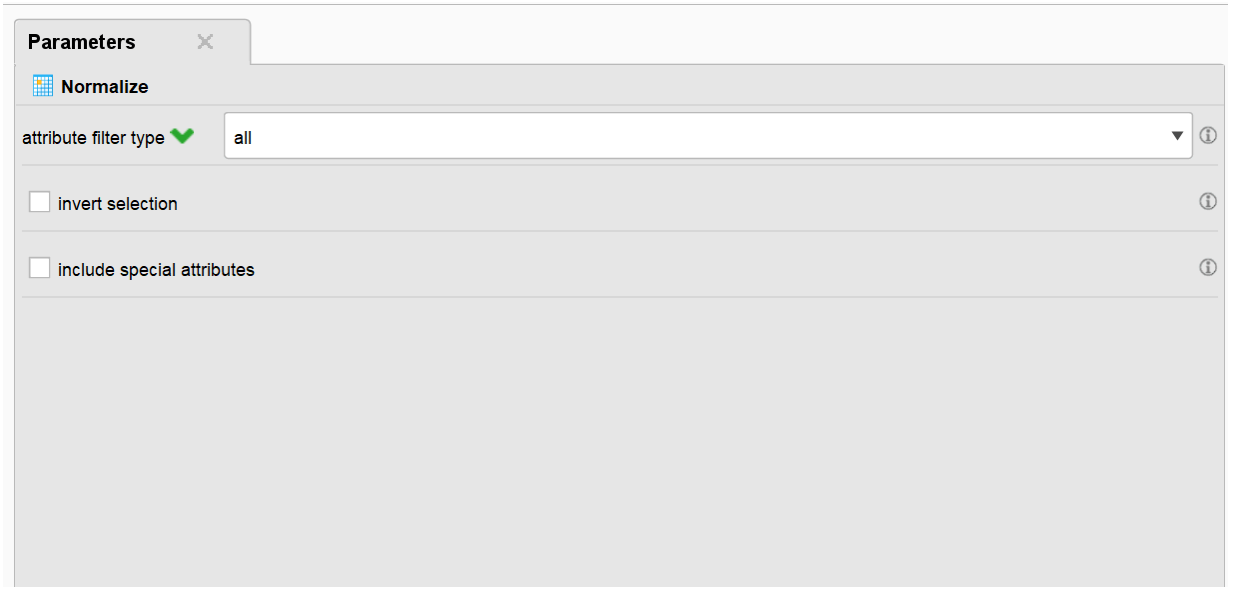


Рисунок 3.3 – настройка блока Normalize

Назначение ролей:

В процессе работы с данными необходимо указать роли для каждого признака. Это делается с помощью блока **"Set Roles"**, блок позволяет настроить, какой атрибут будет являться целевым (label), а какие будут использоваться для обучения модели.

В нашем случае целевая переменная — это MEDV (медианная стоимость жилья, выраженная в тысячах долларов). Для этого в блоке Set Roles необходимо установить роль "label" для этого признака. Это значит, что модель будет предсказывать значение MEDV на основе остальных признаков.

Все остальные столбцы будут использоваться как признаки для обучения модели. Для них назначается роль "regular", что указывает на их использование в качестве независимых переменных для построения модели. Настройка блока показана на рисунке 3.4

Рисунок 3.4 – настройки для Set roles

Разделение данных:

Для разделения данных на обучающую и тестовую выборки используется блок "Split Data". Этот блок позволяет задать пропорцию данных, которые будут использованы для обучения модели и для её тестирования.

В настройках блока Split Data в разделе "partitions" укажите параметр "ratio" равным 0.8. Это значит, что 80% данных будет использовано для

обучения модели, а оставшиеся 20% — для тестирования модели. Настройки блока представлены на рисунке 3.5.

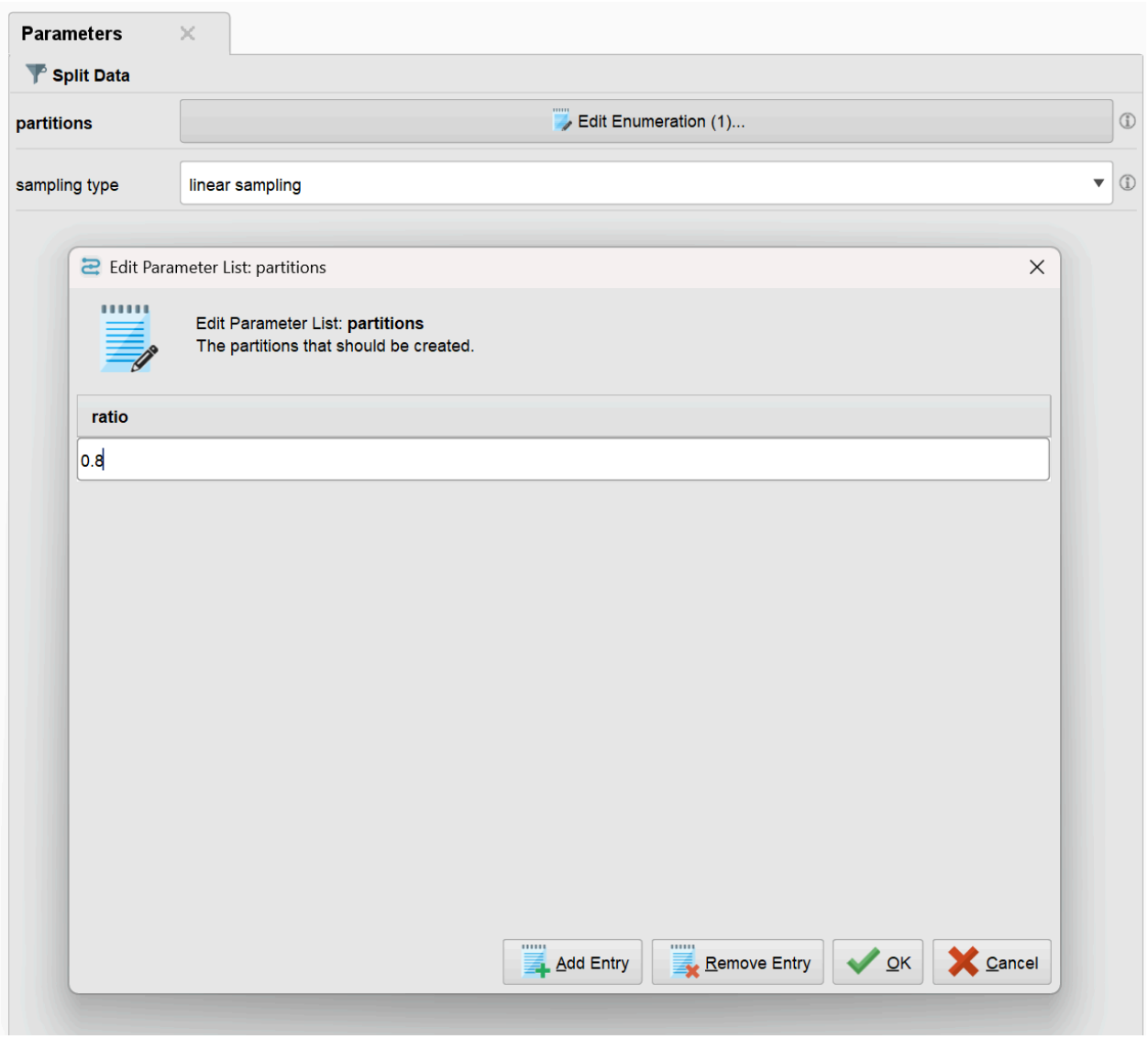


Рисунок 3.5 – настройки для Split Data

Линейная регрессия:

Для построения модели линейной регрессии используется блок **"Linear Regression"**. Этот блок позволяет построить модель, которая будет предсказывать зависимость целевой переменной (в нашем случае **MEDV**) от остальных признаков.

В параметрах блока Linear Regression настраиваются два основных параметра - **min tolerance** и **ridge**. Min tolerance определяет минимальную величину изменения в процессе обучения модели, при которой алгоритм

будет продолжать работу. Установите значение 0.05. Это будет достаточно для большинства задач, чтобы избежать чрезмерного точного подбора модели.

Ridge используется для регуляризации модели, чтобы избежать переобучения. Установите значение 1.0E-8. Это значение обычно подходит для большинства задач с линейной регрессией, так как оно минимизирует влияние регуляризации.

Финальные настройки блока показаны на рисунке 3.6

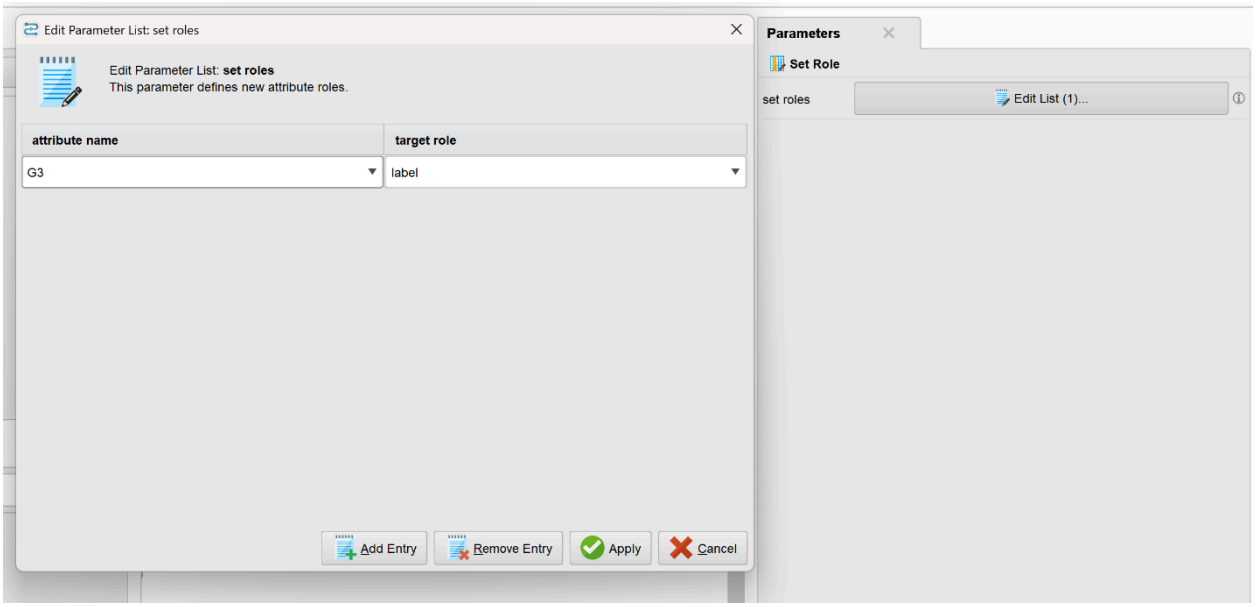


Рисунок 3.6 – настройки для Set roles

Random Forest:

Добавим в проект новый блок **"Random Forest"**, это один из самых мощных методов машинного обучения, который использует ансамбль решающих деревьев для улучшения предсказаний. Он поможет точно спрогнозировать MEDV (цену недвижимости) на основе входных признаков.

В настройках блока Random Forest следующие ключевые параметры:

- **number of trees:** Укажите количество деревьев в ансамбле. В нашем случае устанавливаем 100 деревьев, что является оптимальным значением для большинства задач.
- **criterion:** Установите критерий разделения на "least square". Это стандартный выбор для задач регрессии, так как он минимизирует сумму квадратов ошибок.

- **maximal depth:** Задайте максимальную глубину деревьев как 10. Это предотвращает переобучение и гарантирует, что деревья не будут слишком сложными.
- **apply prepruning:** Установите этот флаг для применения предварительного обрезания деревьев, что поможет улучшить их обобщающую способность и уменьшить вычислительные затраты.
- **minimal gain:** Установите значение 0.01, что означает минимальный прирост, необходимый для разделения узлов.
- **minimal leaf size:** Установите значение 2, что означает минимальное количество экземпляров в листьях дерева.
- **guess subset ratio:** Оставьте этот параметр включенным, что позволяет Random Forest автоматически выбирать подмножества признаков для построения каждого дерева.

Финальные настройки блока представлены на рисунке 3.7.

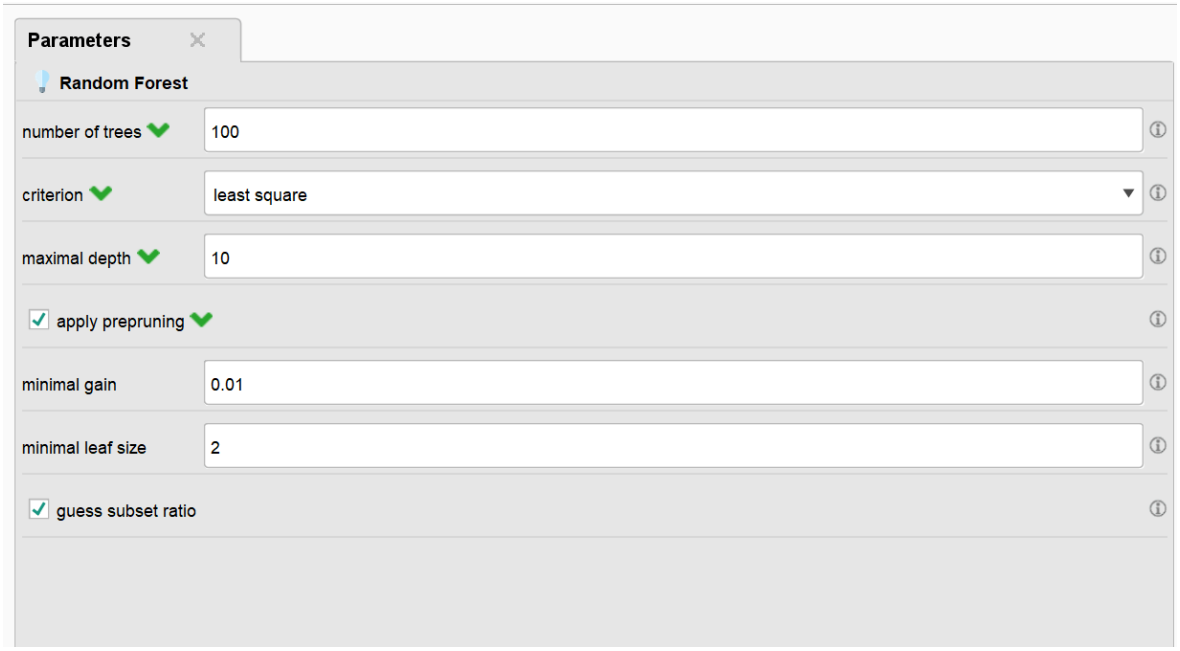


Рисунок 3.7 – параметры для Decision tree

Финальные приготовления к построению дерева:

Для корректного применения обученной модели и последующего анализа её качества необходимо добавить несколько ключевых операторов: **"Apply Model"**, **"Set Role"** и **"Performance"**.

Оператор "Apply Model" используется для тестирования построенной модели на ранее выделенной тестовой выборке. Этот оператор принимает два входных потока данных: обученную модель, созданную с помощью алгоритма линейной регрессии или случайного леса, и тестовый набор данных, полученный из блока "Split Data", на который еще не было наложено предсказание. В результате работы этого оператора к тестовому набору данных добавляется новый столбец, содержащий предсказанные моделью значения целевой переменной (в данном случае MEDV).

После применения модели необходимо правильно задать атрибуту MEDV роль label для дальнейшего анализа. Это выполняется с помощью оператора "Set Role". Добавление этого оператора необходимо, так как аналитические блоки в RapidMiner требуют явного указания целевой переменной, иначе они не смогут правильно интерпретировать данные

Для оценки точности модели был использован оператор "Performance", предназначенный для анализа регрессионных моделей. В параметрах блока были выбраны следующие ключевые метрики оценки:

Root Mean Squared Error (RMSE) — среднеквадратичная ошибка, измеряющая среднее отклонение предсказанных значений от реальных. Это одна из наиболее распространенных метрик для оценки качества регрессионных моделей.

Correlation — коэффициент корреляции, показывающий степень линейной зависимости между предсказанными и фактическими значениями целевой переменной. Это позволяет понять, насколько хорошо модель улавливает взаимосвязь между признаками и целевой переменной.

Все три оператора были последовательно соединены таким образом, чтобы результаты предсказания модели могли быть корректно интерпретированы и проанализированы. Настройки оператора "Performance" и корректное подключение этих блоков представлены на рисунке 3.8

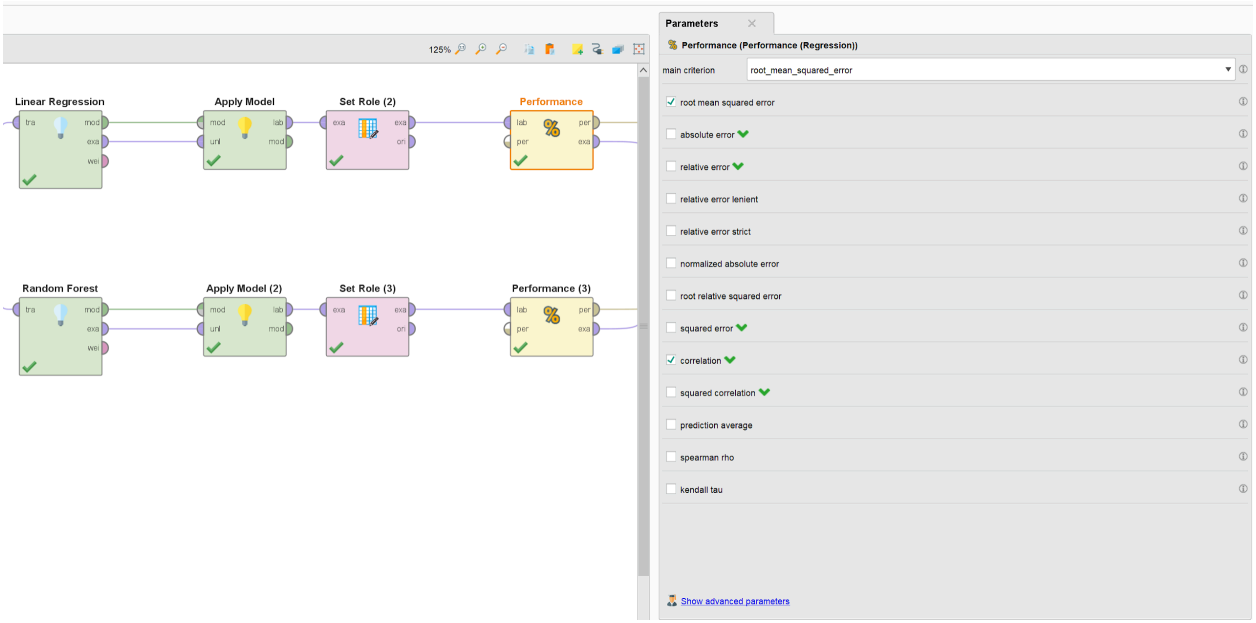


Рисунок 3.8 – Часть проекта и настройка Performance

Оператор Correlation Matrix:

Для дополнительного анализа взаимосвязи между переменными в наборе данных был использован оператор **"Correlation Matrix"**. Этот оператор позволяет вычислить коэффициенты корреляции между всеми числовыми атрибутами. Корреляционный анализ помогает выявить, какие признаки наиболее сильно связаны с целевой переменной, а также оценить возможные мультиколлинеарности между входными переменными.

Настройки оператора

В настройках блока **"Correlation Matrix"** были заданы следующие параметры:

- **Type**: *include attributes* — оператор анализирует только числовые признаки, не учитывая специальные атрибуты (например, ID или категориальные переменные).
- **Attribute filter type**: *all attributes* — корреляция вычисляется для всех числовых атрибутов набора данных.
- **Normalize weights**: *включено* — нормализация весов обеспечивает более корректную интерпретацию корреляций между переменными.
- **Squared correlation**: *выключено* — коэффициенты корреляции отображаются в стандартной форме без возведения в квадрат.

Финальный вид схемы с новым блоком представлен на рисунке 3.9

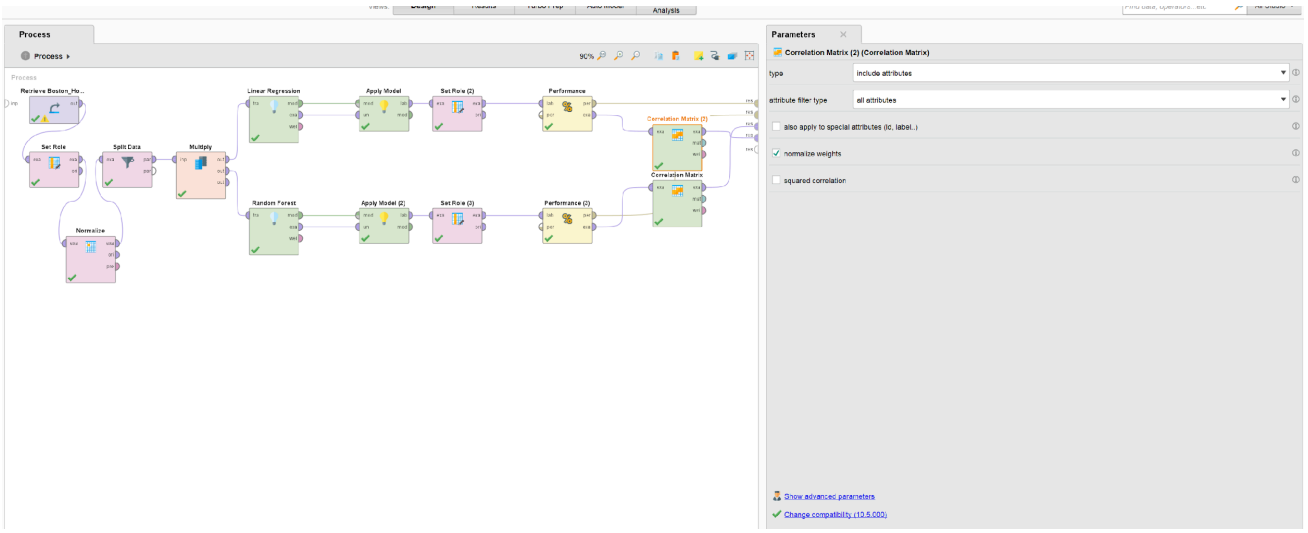


Рисунок 3.9 – Финальный вид схемы

После запуска процесса мы получили несколько результатов, представленных на рисунке 3.10. В частности, были созданы **"ExampleSet (Set Role (2))"** и **"ExampleSet (Set Role (3))"** в которых можно увидеть предсказанные значения prediction (MEDV) для обеих моделей. Значения представлены на рисунках 3.11 и 3.12 соответственно.

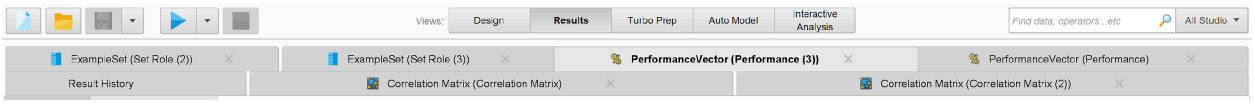


Рисунок 3.10 – Полученные результаты

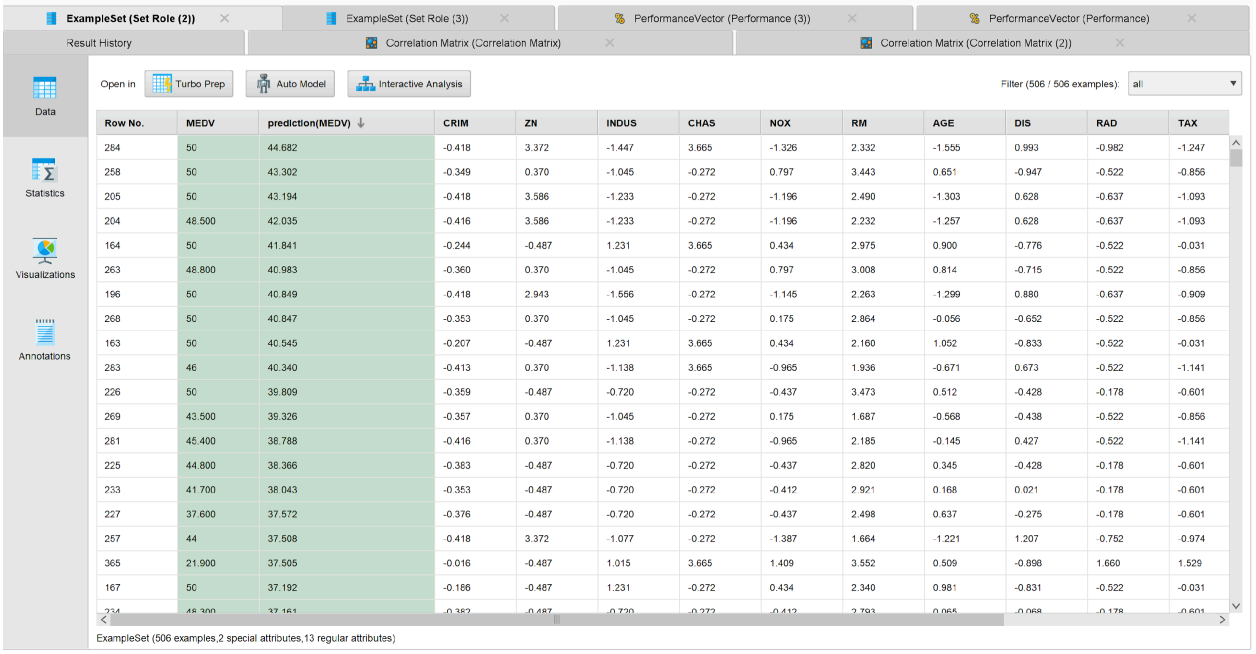


Рисунок 3.11 – ExampleSet (Set role (2))

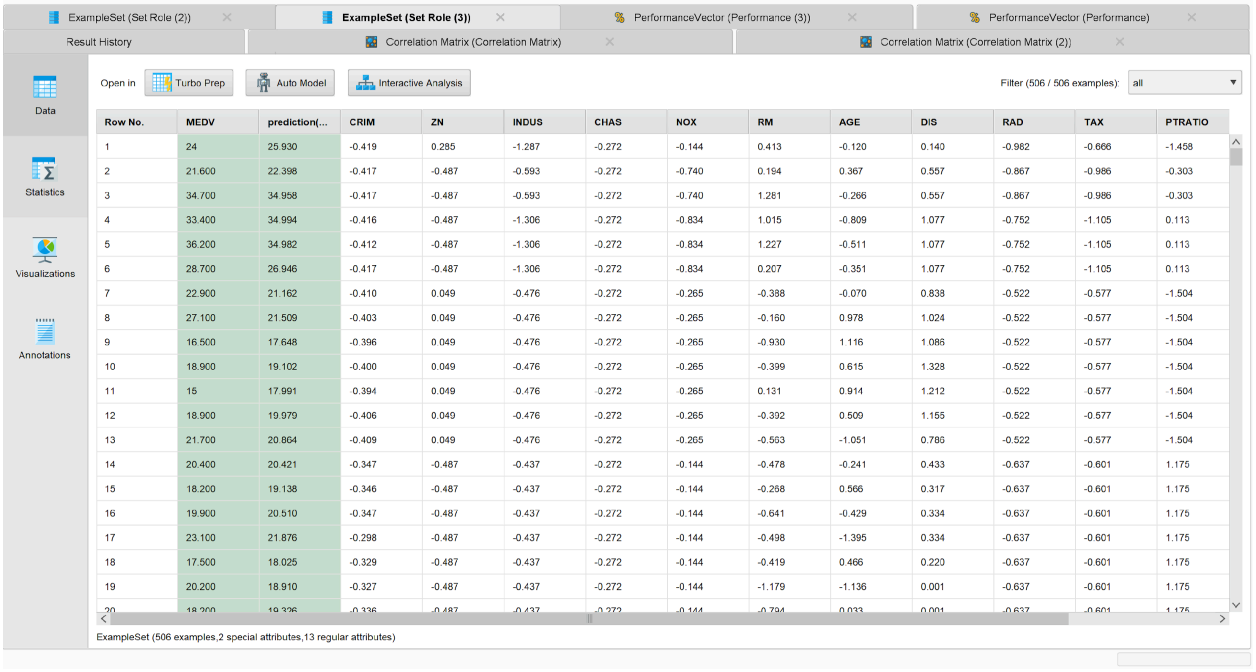


Рисунок 3.12 – ExampleSet (Set role (3))

Интерпретация результатов:

После выполнения модели линейной регрессии были получены следующие ключевые результаты, представленные на рисунках 3.13 и 3.14:

Root Mean Squared Error (RMSE): Значение 4.679 ± 0.000 (рисунок 3.13). Это показатель среднеквадратичной ошибки, который измеряет среднее отклонение предсказанных значений от реальных. Чем меньше это значение, тем точнее модель. В данном случае, низкий RMSE подтверждает, что модель хорошо предсказывает результаты, минимизируя ошибки предсказаний.

Коэффициент корреляции: 0.861 (рисунок 3.14). Это показатель степени линейной зависимости между предсказанными и фактическими значениями. Значение 0.861 указывает на сильную положительную корреляцию, что свидетельствует о высокой точности предсказаний модели. Чем ближе коэффициент корреляции к 1, тем сильнее зависимость, и в данном случае модель подтверждает свою эффективность в прогнозировании.

Таким образом, полученные результаты для модели линейной регрессии демонстрируют её высокую точность, как по методу измерения ошибок (RMSE), так и по степени линейной зависимости между предсказаниями и реальными значениями (коэффициент корреляции)..

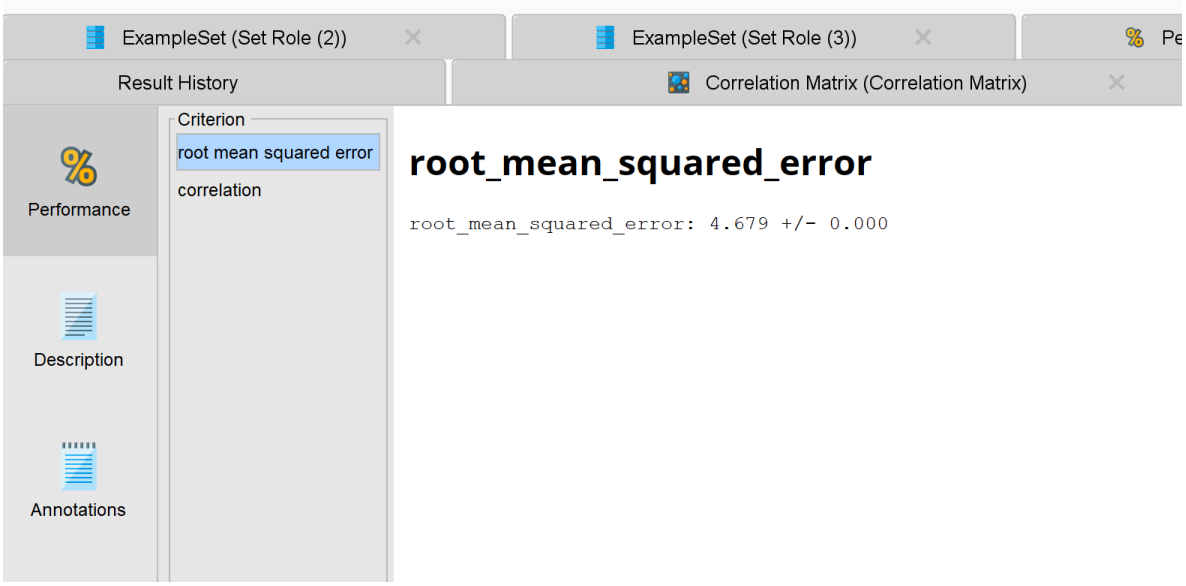


Рисунок 3.13 – RMSE для Liner Regression

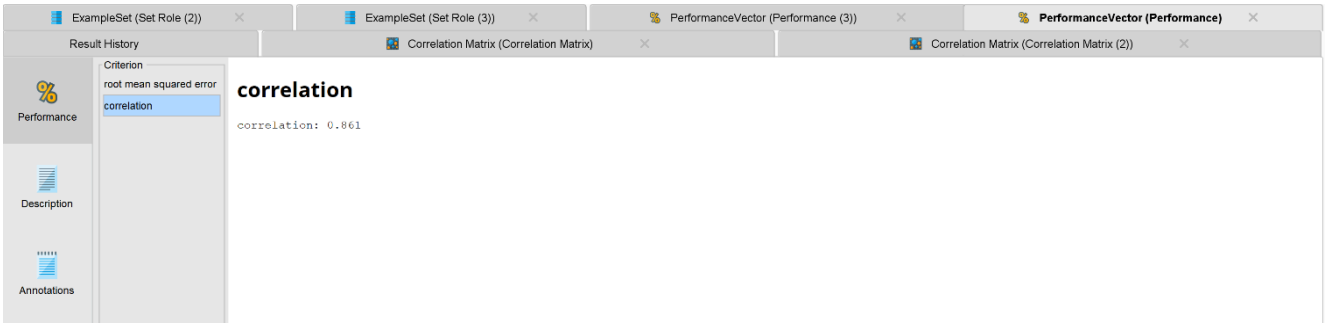


Рисунок 3.14 – Коэффициент корреляции для Liner Regression

Теперь, перейдем к результатам для модели Random Forest. После применения этой модели на тестовой выборке были получены следующие значения:

Среднеквадратичная ошибка (RMSE) для модели Random Forest составила 1.931 (рисунок 3.15), что также свидетельствует о высоком качестве предсказаний. Это значение меньше по сравнению с результатом для линейной регрессии, что подтверждает эффективность модели Random Forest в решении данной задачи.

Коэффициент корреляции между предсказанными значениями и реальными значениями целевого признака MEDV составил 0.981 (рисунок 3.16). Этот результат демонстрирует очень сильную положительную линейную зависимость, что указывает на отличную способность модели предсказывать итоговые значения.

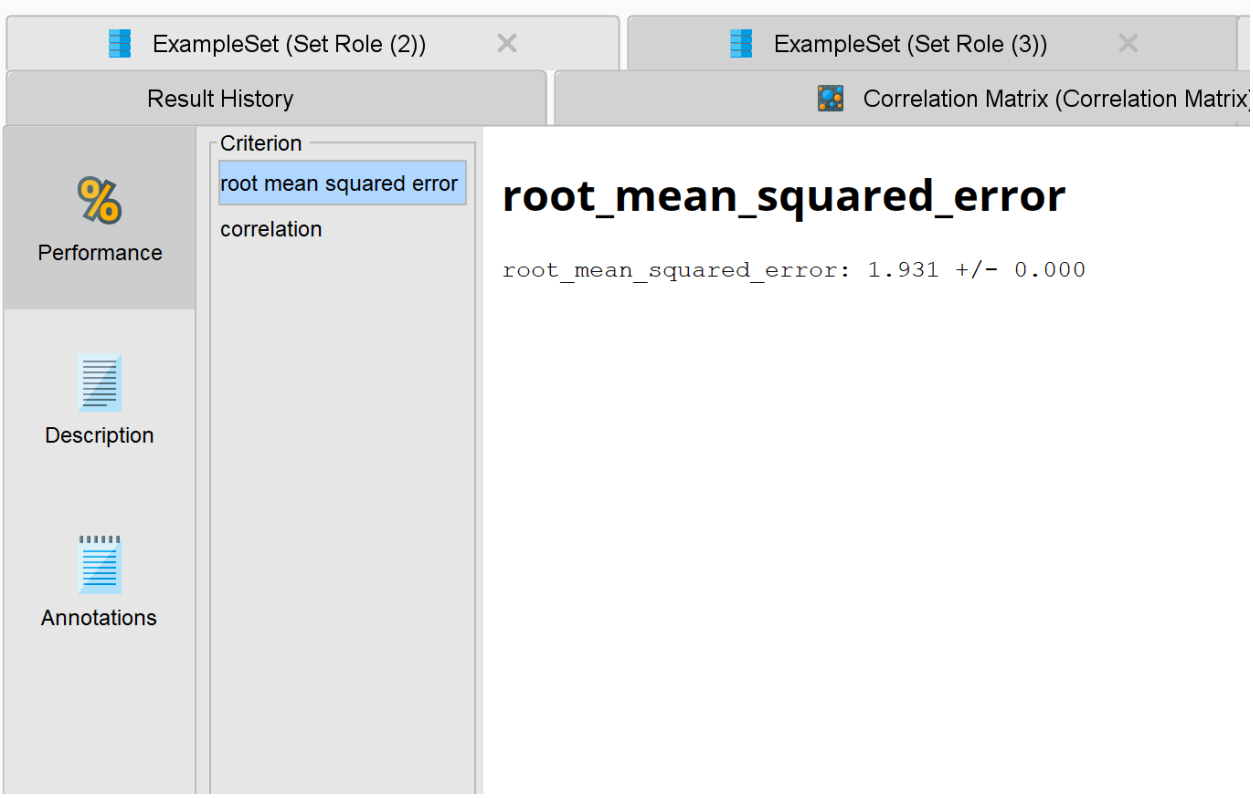


Рисунок 3.15 – RMSE для Random Forest

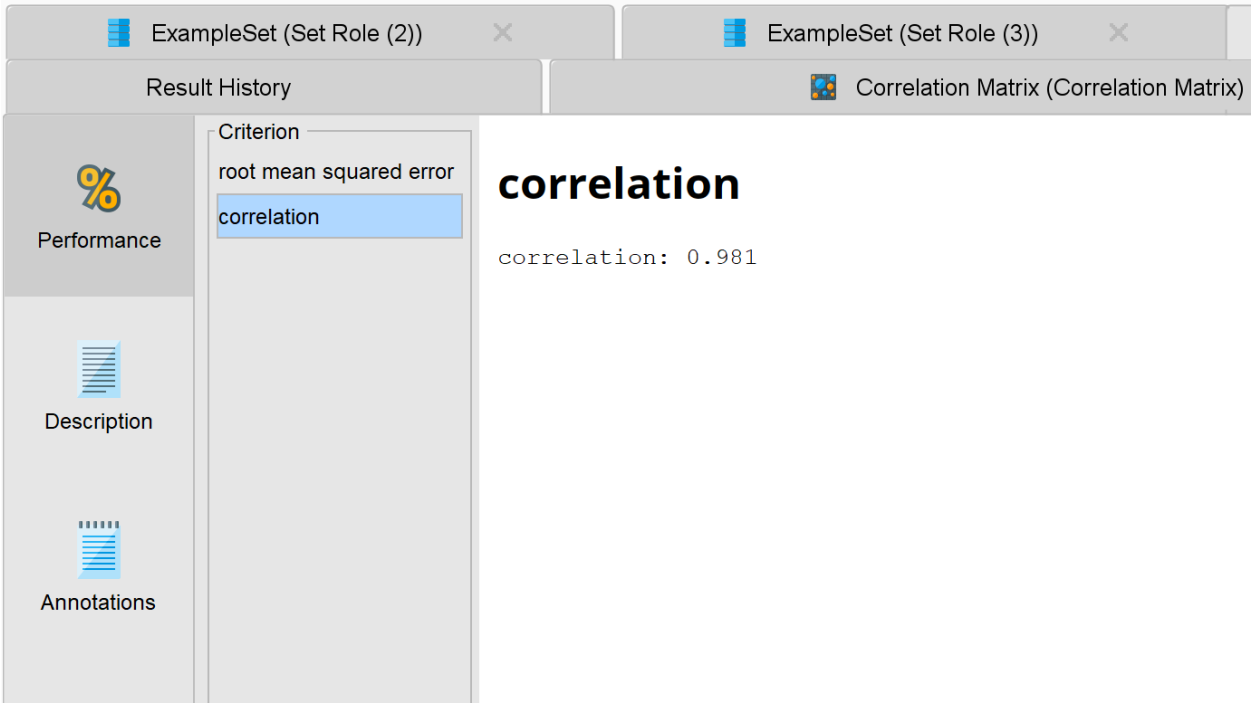


Рисунок 3.14 – Коэффициент корреляции для Random Forest

Заключительным этапом работы стало построение корреляционной матрицы, представленной на рисунке 3.17. Корреляционная матрица позволяет оценить взаимосвязи между всеми признаками в наборе данных и наглядно показывает, какие атрибуты имеют наиболее сильные зависимости

друг с другом. В данной матрице видно, что некоторые признаки имеют высокую степень корреляции, что может указывать на сильное влияние этих признаков на целевую переменную (MEDV).

Например, мы видим высокую корреляцию между TAX и PTRATIO (0.910), что может свидетельствовать о том, что районы с высоким уровнем налогообложения также имеют более высокие коэффициенты ученического состава в школах. Также сильная корреляция наблюдается между INDUS и NOX (0.764), что указывает на связь между уровнем загрязнения воздуха и концентрацией промышленности в разных районах.

Отдельно стоит отметить корреляции между AGE и DIS (-0.748) и RM и LSTAT (-0.742), что может означать, что старые дома (AGE) расположены на больших расстояниях от рабочих центров Бостона (DIS), а также, что высокое количество бедных семей (LSTAT) связано с меньшим количеством комнат в домах (RM). Эти корреляции помогают выявить важные зависимости, которые могут повлиять на результаты модели.

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

Рисунок 3.17 – корреляционная матрица

4. Приобретаемые навыки

1. Работа с интерфейсом RapidMiner Studio – освоение инструментов для построения процессов, анализа данных и настройки операторов.

2. Построение модели машинного обучения – использование алгоритма Random Forest для предсказания целевой переменной.
3. Применение модели к тестовым данным – освоение оператора Apply Model и корректная интерпретация полученных прогнозов.
4. Оценка качества модели – вычисление метрик точности, таких как Root Mean Squared Error (RMSE), и анализ предсказательной способности модели. В ходе работы студенты научатся использовать метрики, такие как RMSE, для оценки точности модели и проверки ее эффективности в реальных условиях.
5. Развитие навыков анализа данных – интерпретация полученных результатов, выявление закономерностей и подготовка отчёта по итогам лабораторной работы. Студенты будут развивать аналитические навыки, интерпретируя результаты работы моделей и формулируя выводы.

5. Обобщенная задача для выполнения индивидуального варианта

Практическая работа направлена на анализ и предсказание непрерывной целевой переменной на предложенном в индивидуальном варианте датасете с числовыми и/или категориальными признаками с помощью методов линейной регрессии и ансамблевого алгоритма Random Forest. В рамках задания необходимо:

1) Подготовка данных:
- Импортировать данные через оператор Read CSV.
- Произвести очистку: обработать пропуски (Replace Missing Values), при необходимости закодировать категориальные признаки (Nominal to Numerical) и удалить нерелевантные столбцы.
- При необходимости нормализовать или стандартизовать числовые признаки.

2) Разделение выборки:
- Разделить данные на обучающую и тестовую части (через Split Data или Cross Validation) в соотношении примерно 70–80% на обучение и 20–30% на тест.

3) Построение моделей:
- Обучить модель Linear Regression с базовыми настройками (и опционально – с регуляризацией).
- Обучить модель Random Forest Regressor, настроив число деревьев, глубину, минимальный размер листа и другие ключевые гиперпараметры.

4) Оценка качества:
- Применить модели к тестовой выборке через оператор Apply Model.
- Оценить качество предсказаний с помощью метрик RMSE, MAE, R^2 и/или Correlation (через оператор Performance).

5) Сравнительный анализ и интерпретация:

- Сравнить полученные метрики у обеих моделей и определить, какая показала лучшие результаты.
- Проанализировать коэффициенты линейной регрессии и важности признаков в Random Forest (через оператор Weights или Feature Importance).
- Построить график «предсказанное vs реальное» и график распределения остаточных ошибок, чтобы визуально оценить разброс и возможные систематические смещения.

6) Выводы:

- Описать, в каких ситуациях и на каком типе данных линейная модель работает лучше, а где выигрывает Random Forest.

6. Распределение вариантов