Machine Learning Master Project: California Housing Dataset

Citizen Data Science Program – Level 2

Guidelines:

1. Answer each question with a clear and concise explanation of the Operator you used. Explain why you chose that particular operator and how it helps.
2. Outline the steps you followed to arrive at the solution.
3. Provide a screenshot of your process that demonstrated the relevant portion of your answer.
4. Name the document as the {project name - your name} before submitting it.
5. Ensure that you include both your answers word document and RapidMiner process as attachments in a reply to the person who sent you the assessment.
6. Write your name and email at the bottom of this page

Project Description

This project consists of 16 questions related to your process in RapidMiner of the California Housing dataset. The questions are designed to assess your understanding of the key concepts and your ability to apply them in practical scenarios.

You will work on the California Housing dataset using RapidMiner to build a model of housing prices to predict median house values in California. The dataset contains information about housing prices in various districts of California. The dataset includes the following features:

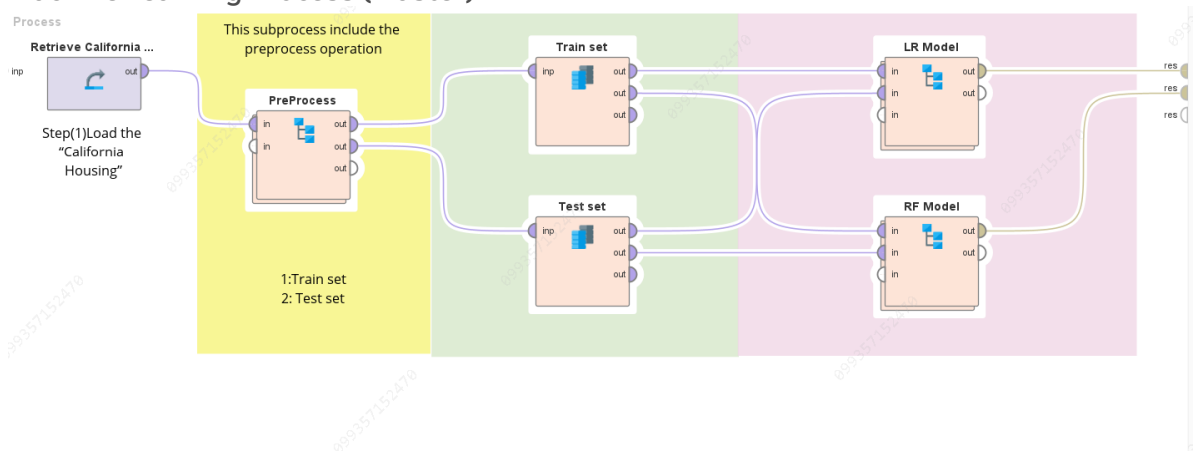
1. longitude: Longitude value for the block in California, USA
2. latitude: Latitude value for the block in California, USA
3. housing_median_age: Median age of the house in the block
4. total_rooms: Count of the total number of rooms (excluding bedrooms) in all houses in the block
5. total_bedrooms: Count of the total number of bedrooms in all houses in the block
6. population: Count of the total number of population in the block
7. households: Count of the total number of households in the block
8. median_income: Median of the total household income of all the houses in the block
9. ocean_proximity: Type of the landscape of the block [Unique Values : 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND']
10. median_house_value: Median of the household prices of all the houses in the block

Name: Abdullah Mohammad Al Talaq

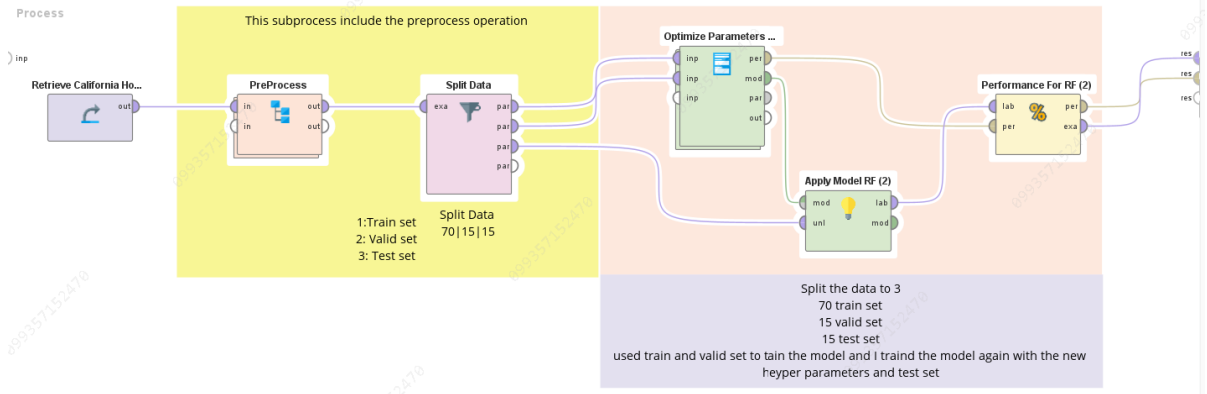
Email: AlTalaqA@sabic.com

Section 0: The project includes three files.

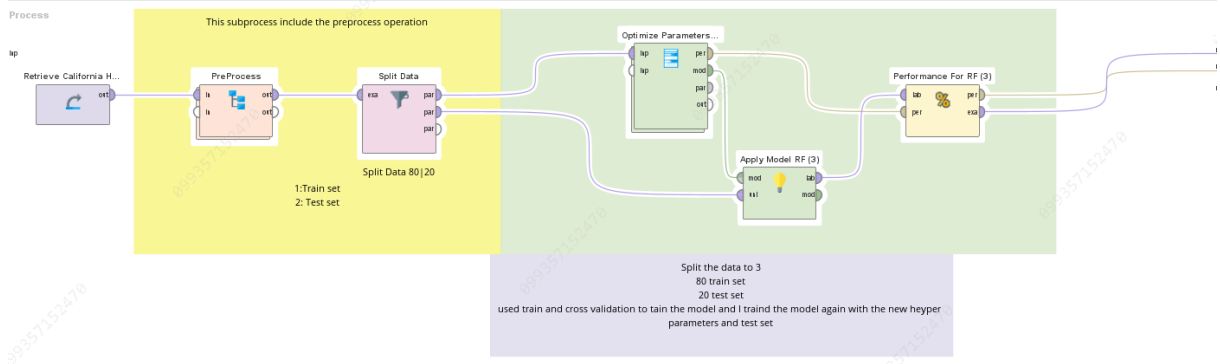
1- Machine Learning Process (Master)



2- Machine Learning Process (Master) Optimizations



3- Machine Learning Process (Master) Optimizations cross-validation



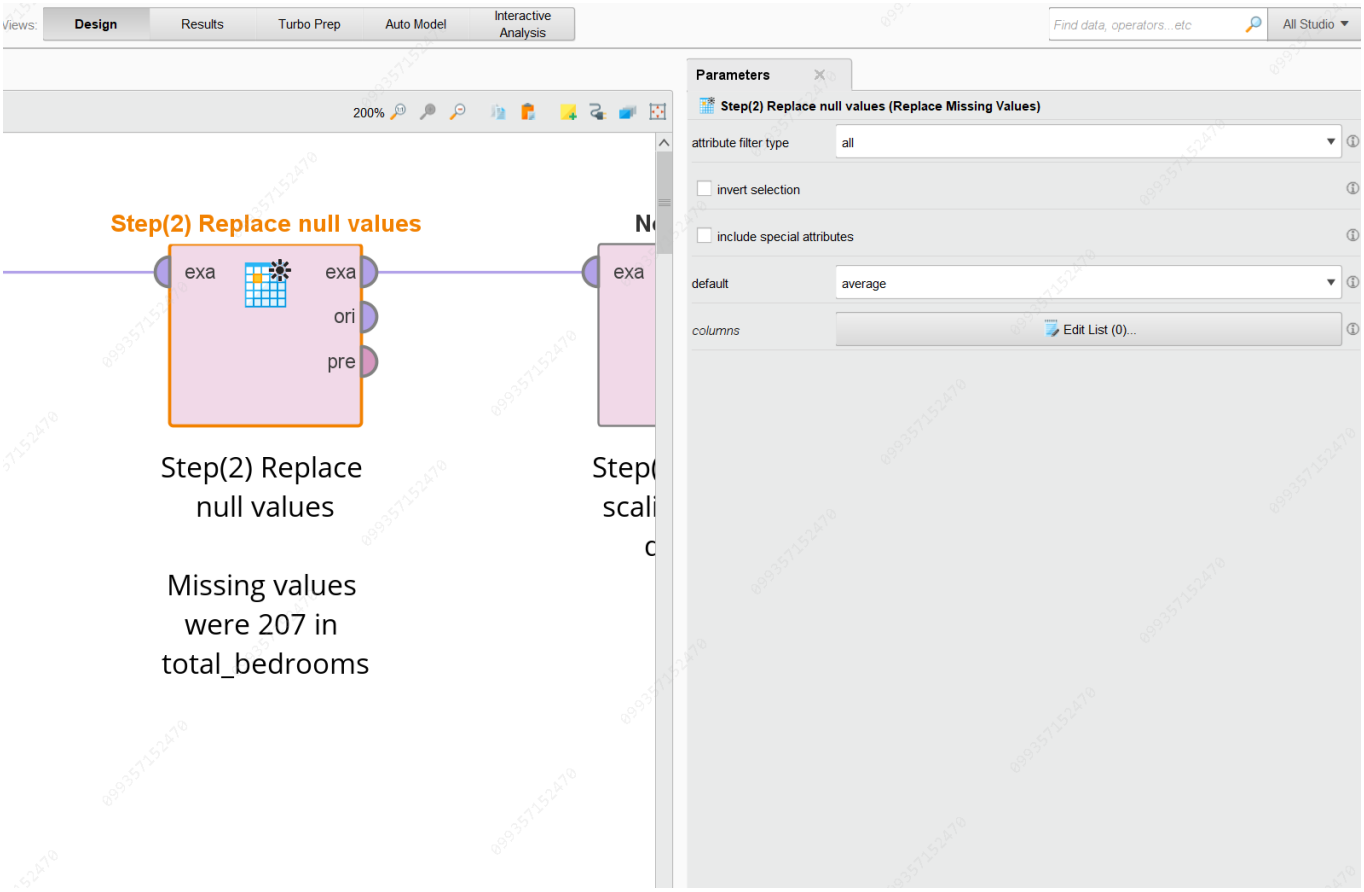
Note there are sub process operations. To check you have to open the process in Ai studio (RapidMiner)

Section 1: Data Exploration & Preprocessing

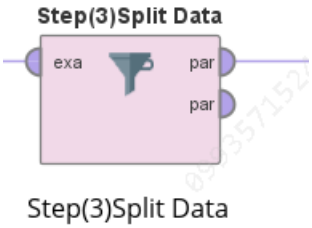
1. Preprocess the dataset by handling the missing values, if any, and. How many missing values are there? And how did you handle them?

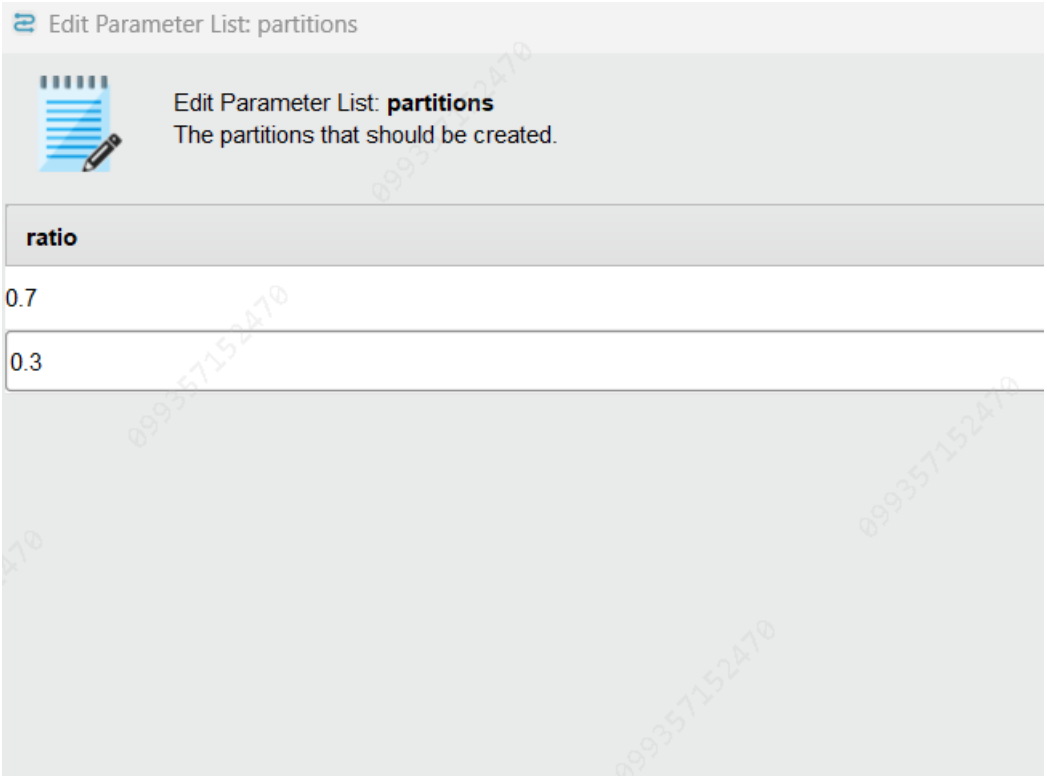
			Min	M...	Average
total_bedrooms	Integer	207	1	6445	537.871

Handling:

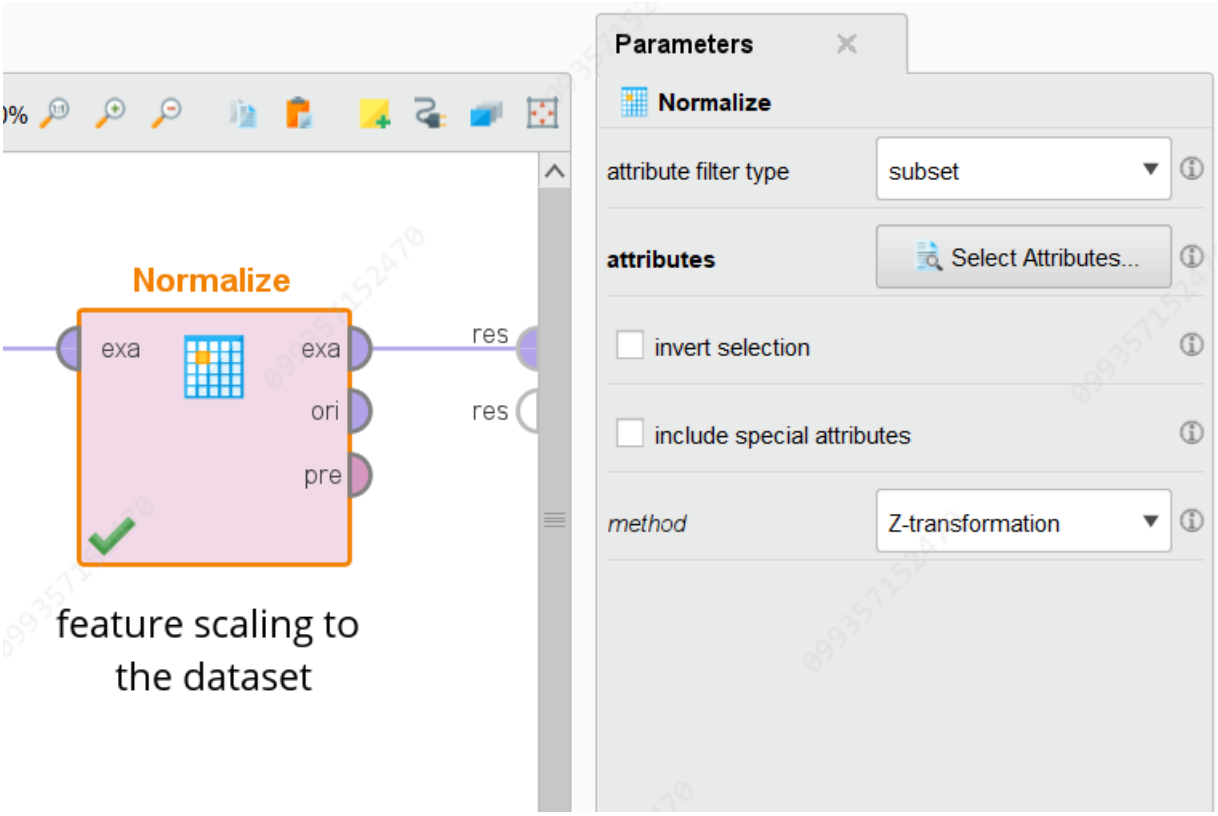


2. Split the dataset into training and testing sets using a 70:30 split ratio.





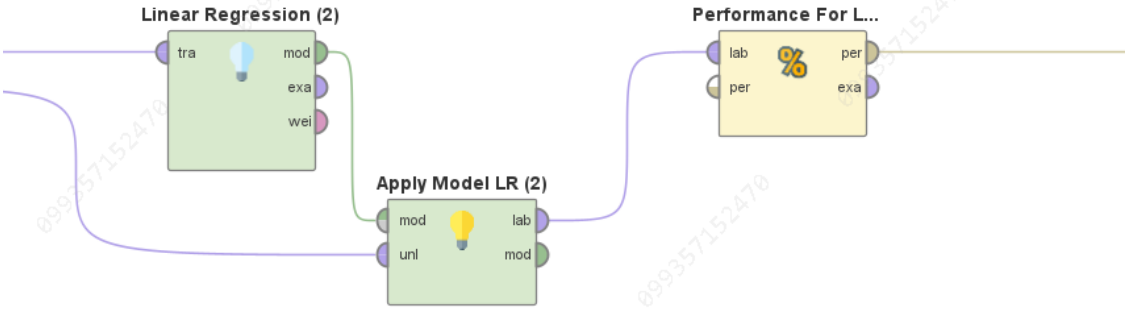
3. Apply feature scaling to the dataset. Which scaling method did you use, and why this step is important?

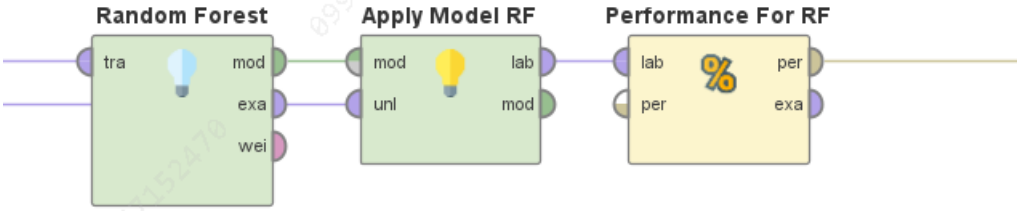


Important process of standardizing data formats to ensure consistency and accuracy. By eliminating redundancies and errors, data normalization makes managing large volumes of information smoother and more efficient.

Section 2: Model Selection & Evaluation

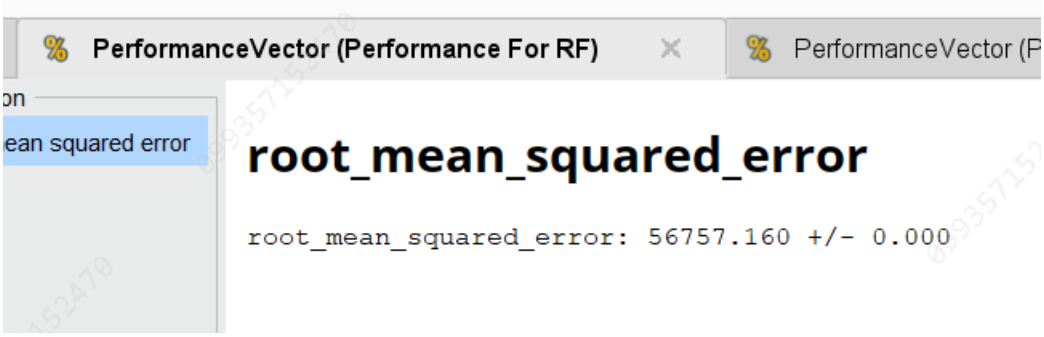
1. Choose two algorithms from the given options: [Linear Regression, Random Forest, or Gradient Boosting Trees]. Justify your choice for each algorithm based on characteristics and suitability for the given dataset.



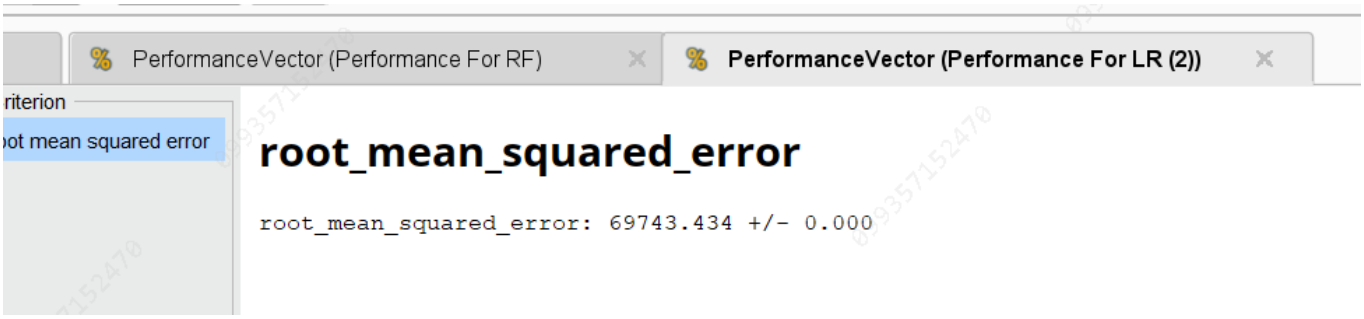


Justify for Linear: outcome variable is a numeric variable, and linear regression has fast and simplicity.
Justify for Random Forest: random forests models are not as sensitive to hyperparameters as other models, so it is easy to get a decent model up and running without having to do too much parameter tuning.

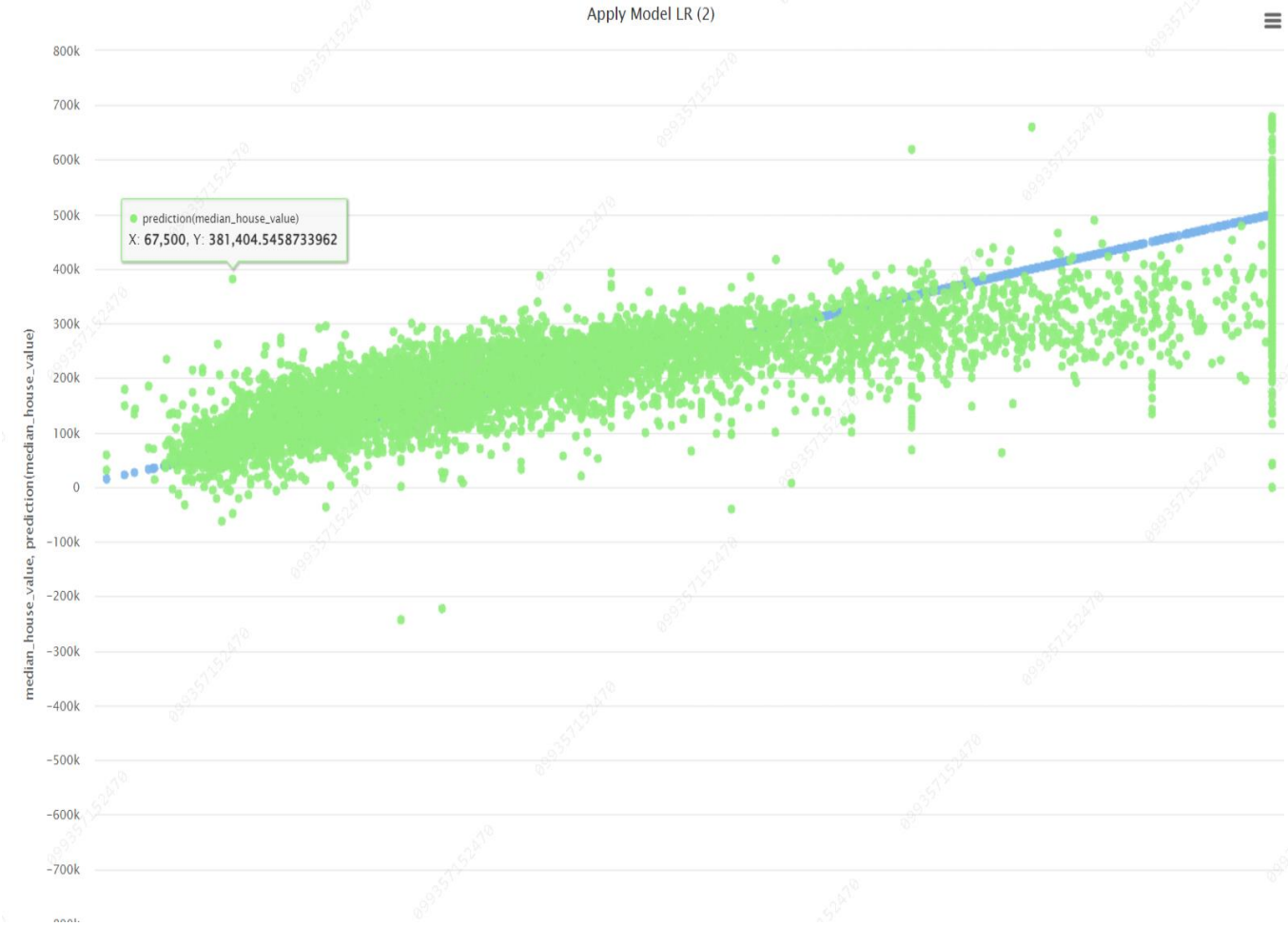
2. Train the first chosen algorithm on the training data to predict housing prices. Evaluate its performance using RMSE on testing data and explain what the RMSE performance metric is?

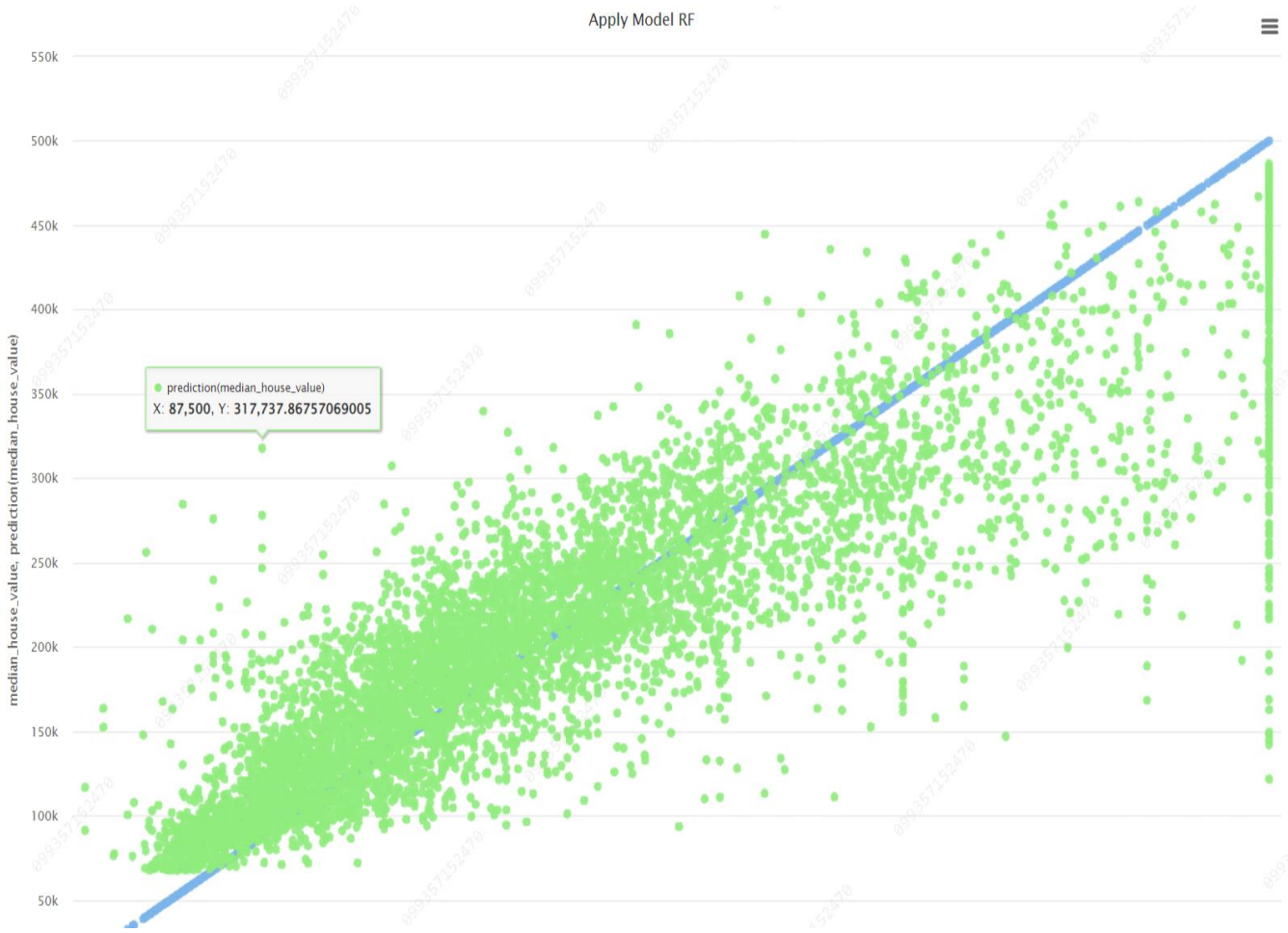


3. Train the second chosen algorithm on the training data and evaluate its performance on testing data (RMSE).



4. Compare the performance of the two chosen algorithms. Which algorithm performs better for predicting housing prices, and why?

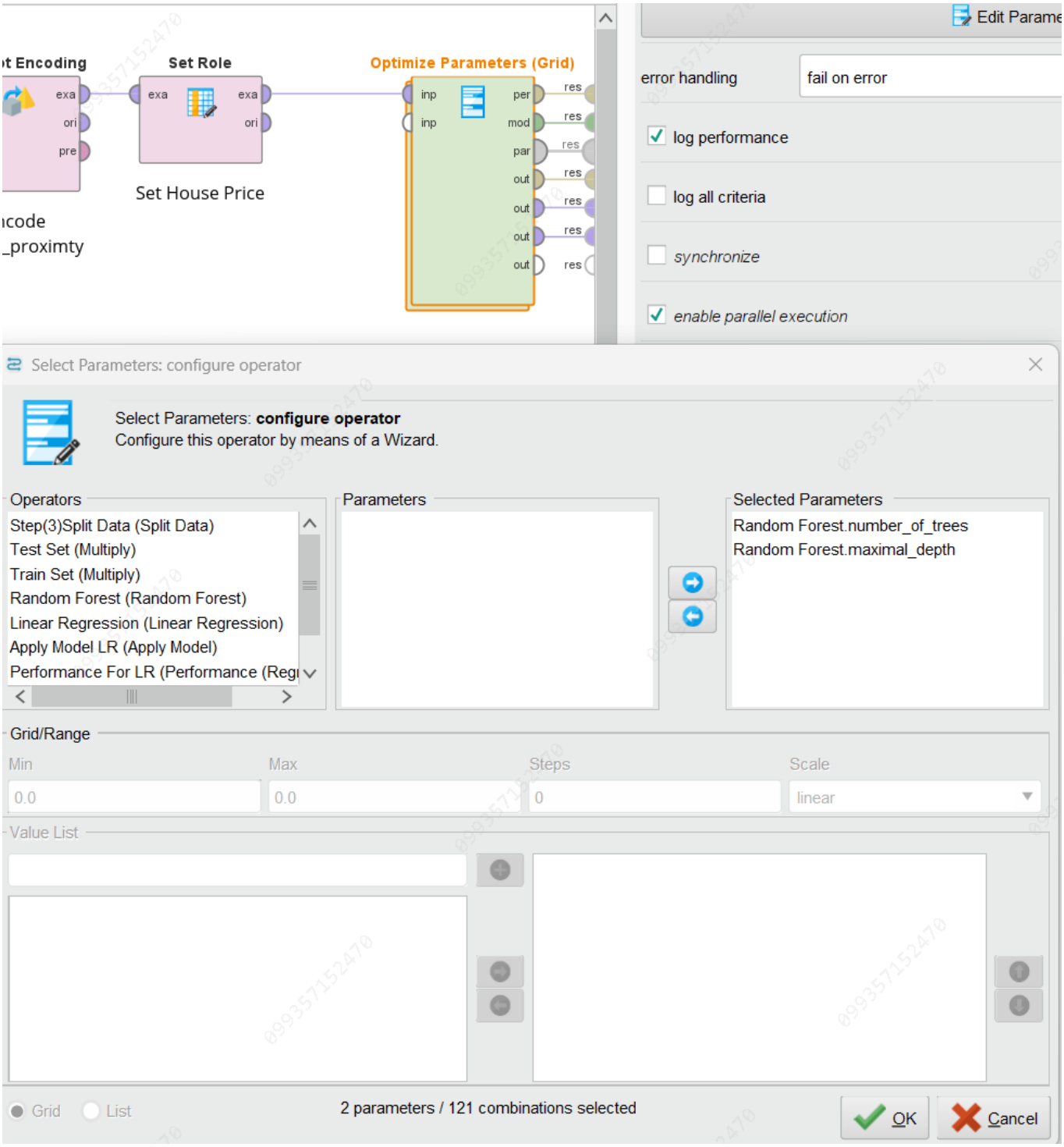




Random forest performs better then Linear regression. RF predictions close to actual values and you barely find outliers while in LR there are some predictions are far from the range of y-axis for actual data. Also, there are prediction values less than zero, which is unlogical.

Section 3: Hyperparameter Tuning & Optimization

1. Apply grid search or random search optimization to tune the hyperparameters of the better performing algorithm. Specify the hyperparameters you selected for optimization and explain their significance in the context of the chosen algorithm.



Number of trees: The main advantage of using number of trees that predictive performance tends to increase as the number of trees increases.
Maximal depth: Advantage of creating trees with a high max depth is that its predictive performance may benefit from doing so. In general, adding more splits to your trees will result in better classification.

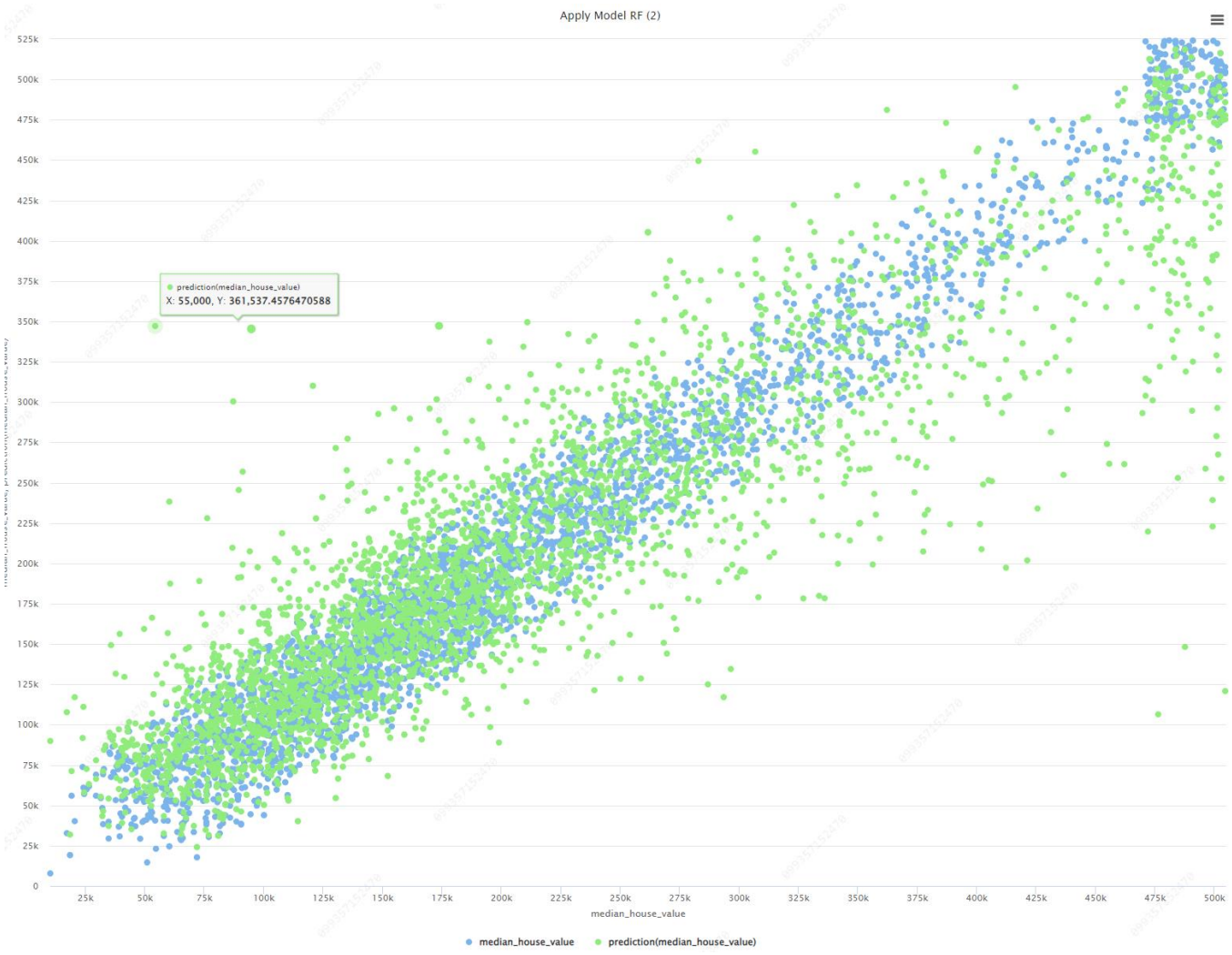
2. Train the optimized model on the training data and evaluate its performance on testing data (RMSE).

PerformanceVector (//Local Repository/Result/Machine Master/Oprimize+CrossValid)

root_mean_squared_error

```
root_mean_squared_error: 49797.427 +/- 0.000
```

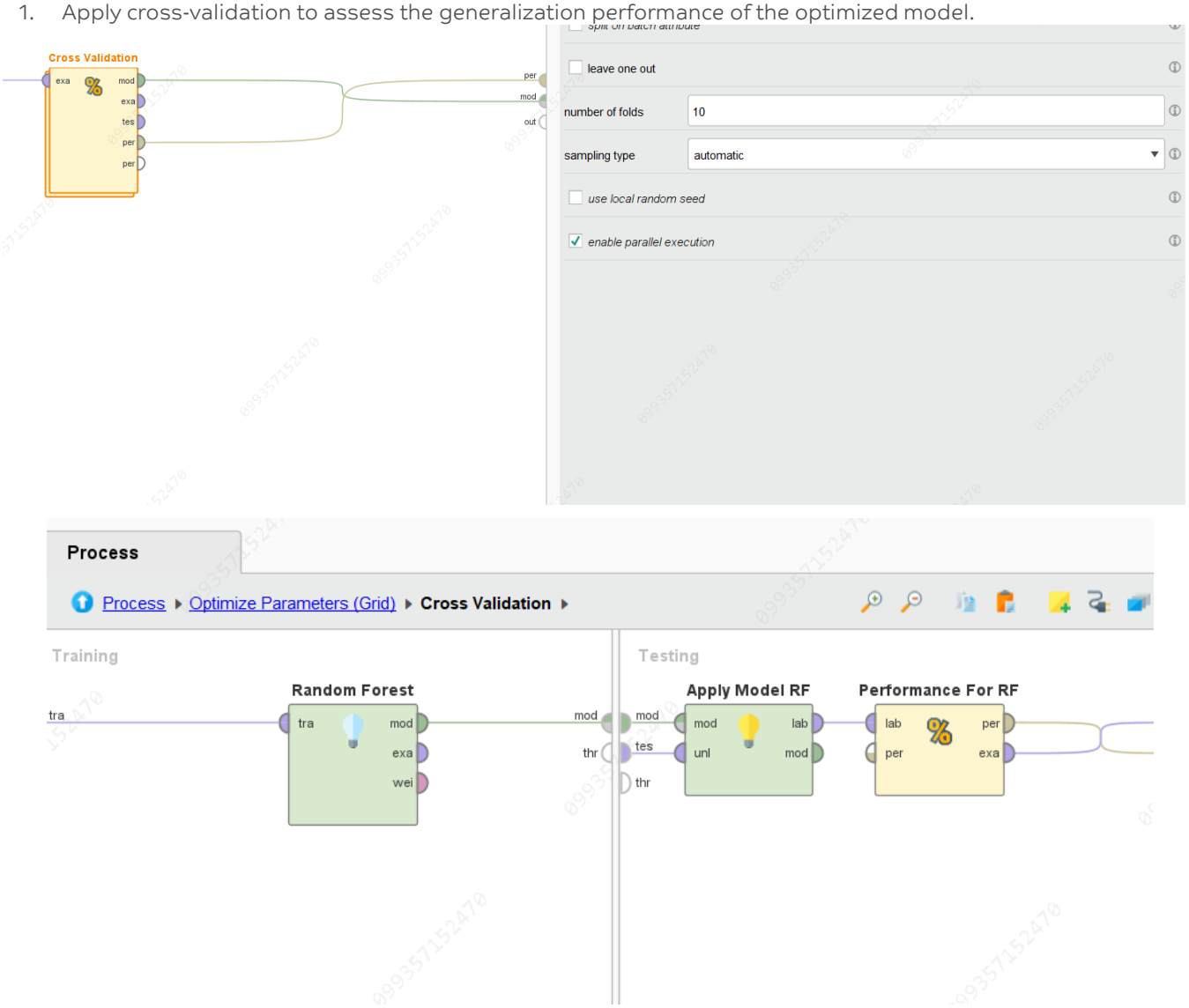
3. Show the performance of the optimized model with the previous model. Has the performance improved after optimization?



The performance becomes better, RMSE decreased.

Section 4: Model Interpretation & Analysis

1. Apply cross-validation to assess the generalization performance of the optimized model.



Section 5: Overfitting & Underfitting Analysis

1. Explain overfitting and underfitting in machine learning.

   overfitting: A model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

   Underfitting: A model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data.

Section 6: Algorithm Strengths, Recommendation, and Implications

1. Summarize your findings and provide recommendation for predicting housing prices based on the provided dataset.

There should be an attribute for the city which the housing is in, having a data on the entirety of California is far more generalized to be modeled. According to google. California population is 39 million, and 485 cities. To make suitable model for such big city like this we need bigger dataset than this.

Feel free to explore and experiment with the dataset and RapidMiner's capabilities beyond the questions mentioned above.