First way to improve the data: Using the feature importances function to get information of the features.

The feature importances function in Random Forest, a popular ensemble learning method used for both classification and regression tasks, provides insights into which features (or variables) contribute most to the prediction accuracy of the model. Random Forest operates by constructing multiple decision trees during training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Feature importance is a way to understand which features have the most influence on the predictions made by the Random Forest model.

This step will helps in understanding which features are most influential in predicting the target variable. This can guide domain understanding and feature engineering. Also, understanding which features are important (and which are not) can guide efforts in feature engineering, such as creating interaction terms between important features or exploring transformations that could make some features more informative.

Second way to improve the data: Using bootstrap to load the model with more data

Bootstrap is a powerful statistical technique used for estimating the distribution of a statistic (like the mean, median, variance) from a set of data by resampling with replacement. It belongs to a broader class of resampling methods used for statistical inference, notably for estimating the sampling distribution of a statistic, assessing its variability, and constructing confidence intervals or hypothesis tests when traditional parametric assumptions cannot be met or when a theoretical distribution is difficult to obtain.

Bootstrap allows for estimating parameters of the population from which the sample was drawn. This is particularly useful when the theoretical distribution of the estimator is complex or unknown. It will also helps in assessing the variability or standard error of a statistic since the spread of the bootstrap distribution gives an indication of the statistic's variability.

Bootstrap can be used to compare statistics (like means or medians) from different samples by examining the overlap between their confidence intervals derived through bootstrapping.

Third way to improve the data: Ensuring Data Quality During Collection

Improve data collection processes to ensure high-quality data from the start. This could involve setting up validation rules to prevent incorrect data entry, designing surveys to minimize biases, or using high-quality sensors in IoT applications.

The data is collected from Boston's official government website and the data is pre checked and cleaned by the website so the data good enough to be used in this project.

By minimizing errors and inconsistencies at the source, there is less need for extensive data cleaning and preprocessing later on. This saves time and resources in the data analysis phase.

High-quality data leads to more reliable insights and better decision-making. When data accurately reflects the real-world phenomena it is intended to represent, the conclusions drawn from it are more trustworthy. In machine learning, the quality of the input data is a key determinant of model performance. High-quality data can lead to more accurate and robust models.

After all three of steps, the accuracy score improved a little.

The best accuracy I got is: 0.9946347769877182

And before the data centric step, the accuracy score is: 0.9943762120232709

The data centric steps did helped to improve the accuracy. A data-centric approach often involves augmenting the dataset with more varied or representative samples. This can help the model generalize better to unseen data, improving its accuracy in real-world applications. Data-centric methods might involve thoughtful feature engineering to highlight relevant information that can help models make more accurate predictions. This includes creating or selecting features that effectively capture the essential aspects of the data.