Module: BI: Data Analytics 03.05.2017

Professor: Pouya Yousefi

Student: Gizem Kaptan (MoTIS 8)

Final Project: Boston Housing Prices Prediction by using Dataiku DSS

Objective

The objective of working with this dataset is to predict the median value of owner occupied homes in the suburbs of Boston, USA.

Dataset

The dataset comes from real estate industry in Boston.

Resource: http://archive.ics.uci.edu/ml/datasets/Housing

Data:

Housing Data.csv

Attribute Information:

1. CRIM: per capita crime rate by town

2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

3. INDUS: proportion of non-retail business acres per town

4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5. NOX: nitric oxides concentration (parts per 10 million)

6. RM: average number of rooms per dwelling

7. AGE: proportion of owner-occupied units built prior to 1940

8. DIS: weighted distances to five Boston employment centres

9. RAD: index of accessibility to radial highways

10. TAX: full-value property-tax rate per $10,000

11. PTRATIO: pupil-teacher ratio by town

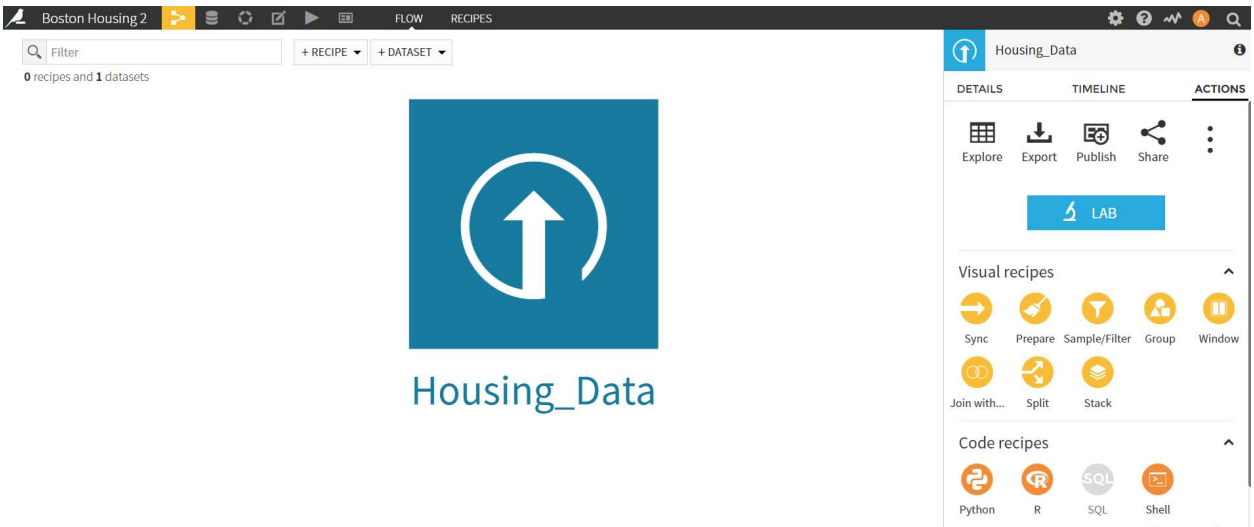
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT: % lower status of the population

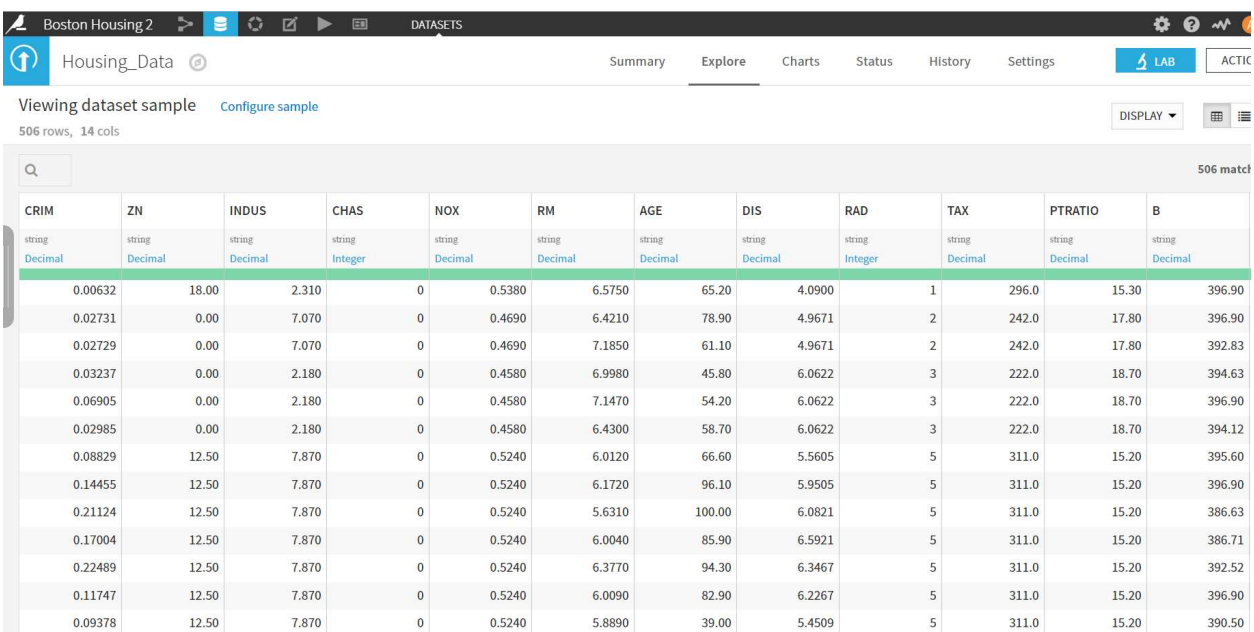
14. MEDV: Median value of owner-occupied homes in $1000's

1. Data Import and Preparation

First step is to import the data in DSS. After importing the data, the flow will look like below.



The data will appear in DSS as shown below.



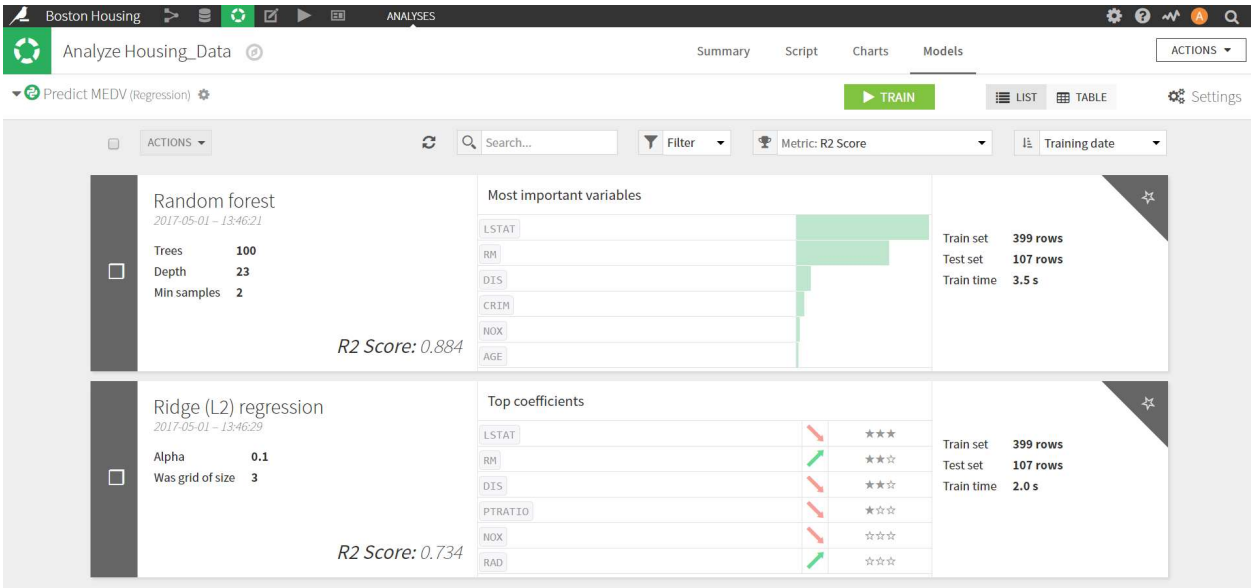
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B
string Decimal	string Decimal	string Decimal	string Integer	string Decimal	string Decimal	string Decimal	string Decimal	string Integer	string Decimal	string Decimal	string Decimal
0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90
0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90
0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83
0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63
0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90
0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12
0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60
0.14455	12.50	7.870	0	0.5240	6.1720	96.10	5.9505	5	311.0	15.20	396.90
0.21124	12.50	7.870	0	0.5240	5.6310	100.00	6.0821	5	311.0	15.20	386.63
0.17004	12.50	7.870	0	0.5240	6.0040	85.90	6.5921	5	311.0	15.20	386.71
0.22489	12.50	7.870	0	0.5240	6.3770	94.30	6.3467	5	311.0	15.20	392.52
0.11747	12.50	7.870	0	0.5240	6.0090	82.90	6.2267	5	311.0	15.20	396.90
0.09378	12.50	7.870	0	0.5240	5.8890	39.00	5.4509	5	311.0	15.20	390.50

Since this data is clean and ready to analyze and train prediction models, I didn't carry out any cleaning, preparation or enriching steps or I didn't apply any recipes before starting with creating predictions.

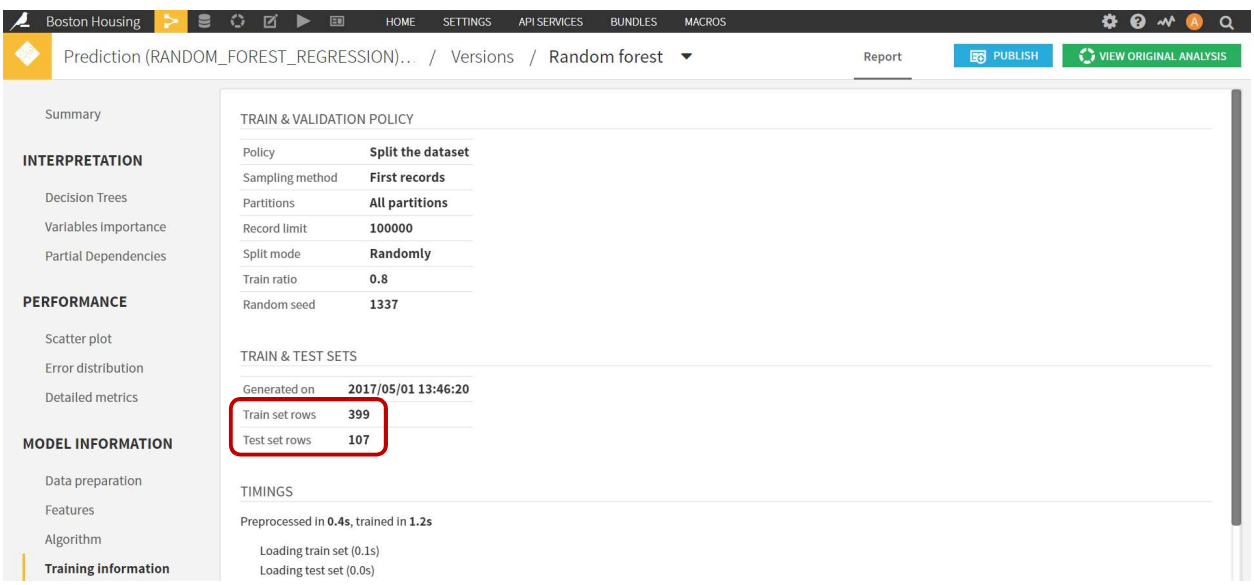
2. Creating Prediction Models

In order to create prediction models, we open an analysis on the dataset and create a first prediction model with the price (MEDV) as target feature.

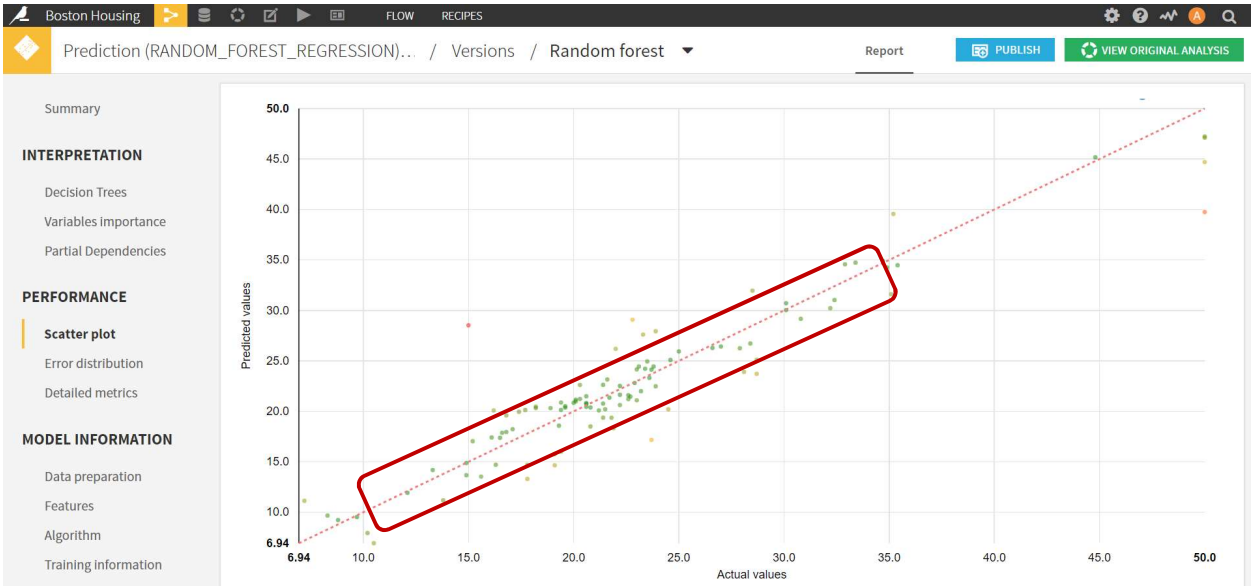
DSS trained automatically two models on the dataset and display them in a list as below.



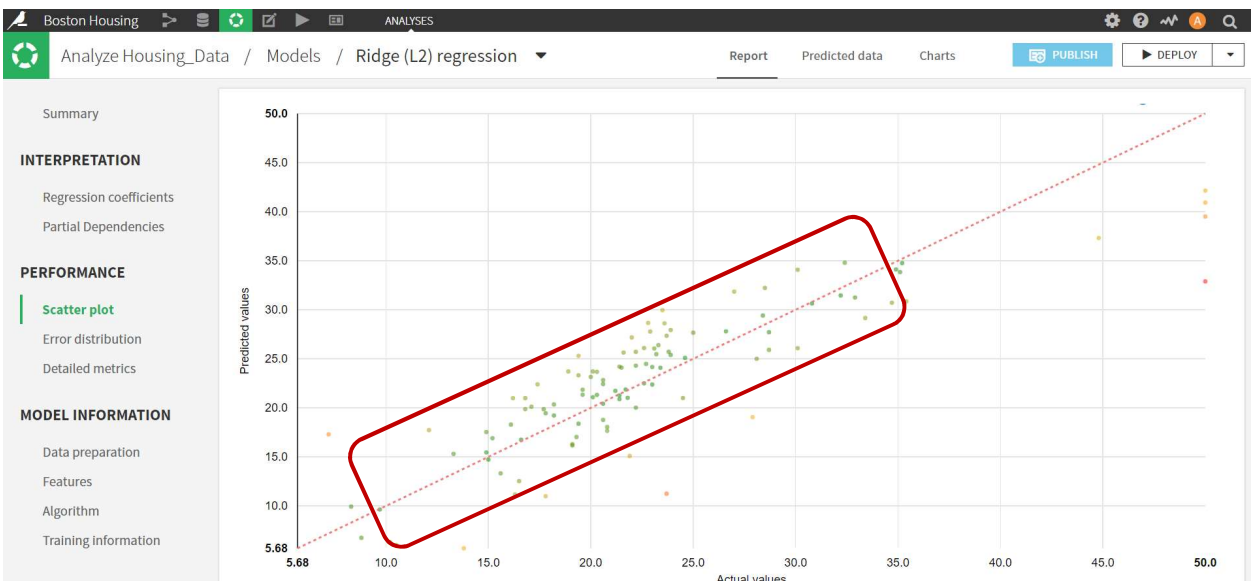
DSS begins to train a model by splitting the dataset into train and test datasets. I clicked on the Random Forest model and in **training information** tab, the train and test dataset information can be found.



In **scatter plot** tab the graph displays the actual value of the target on the x-axis, and the value of the target predicted by the model on the y-axis. If the model was perfect, all points would be on the line.

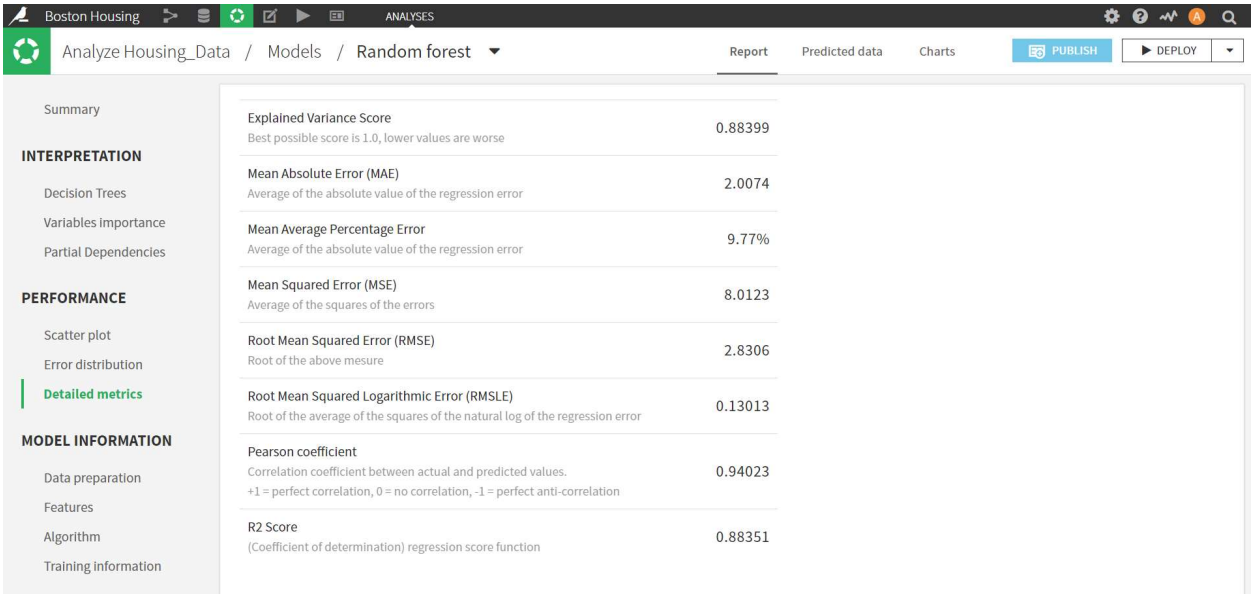


Scatter plot of the Ridge (L2) regression model:

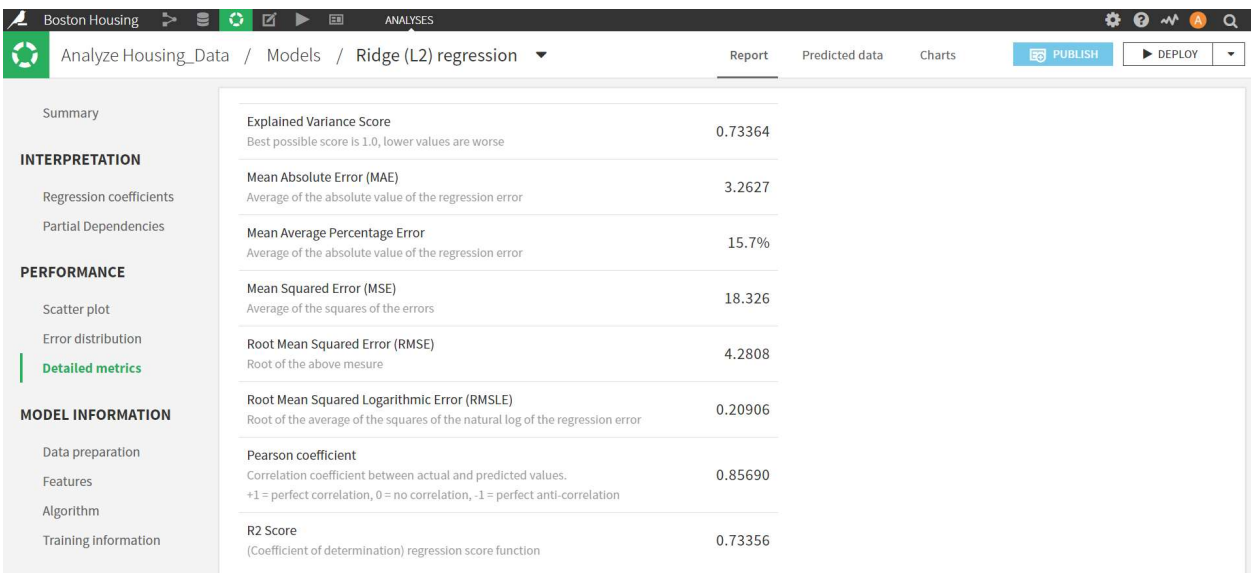


If we compare the prediction results of the 2 models just by looking at the graphs, we can recognize that, the majority of the plots in the Random Forest model are located very close to the line, while the plots in the Ridge regression model are spread around the line with longer distances compared to the first model.

Another tool to measure the performance of the model is using standards statistical scores, which are provided by DSS. In **detailed metrics** tab the scores of the corresponding model according to all defined statistical metrics can be found.

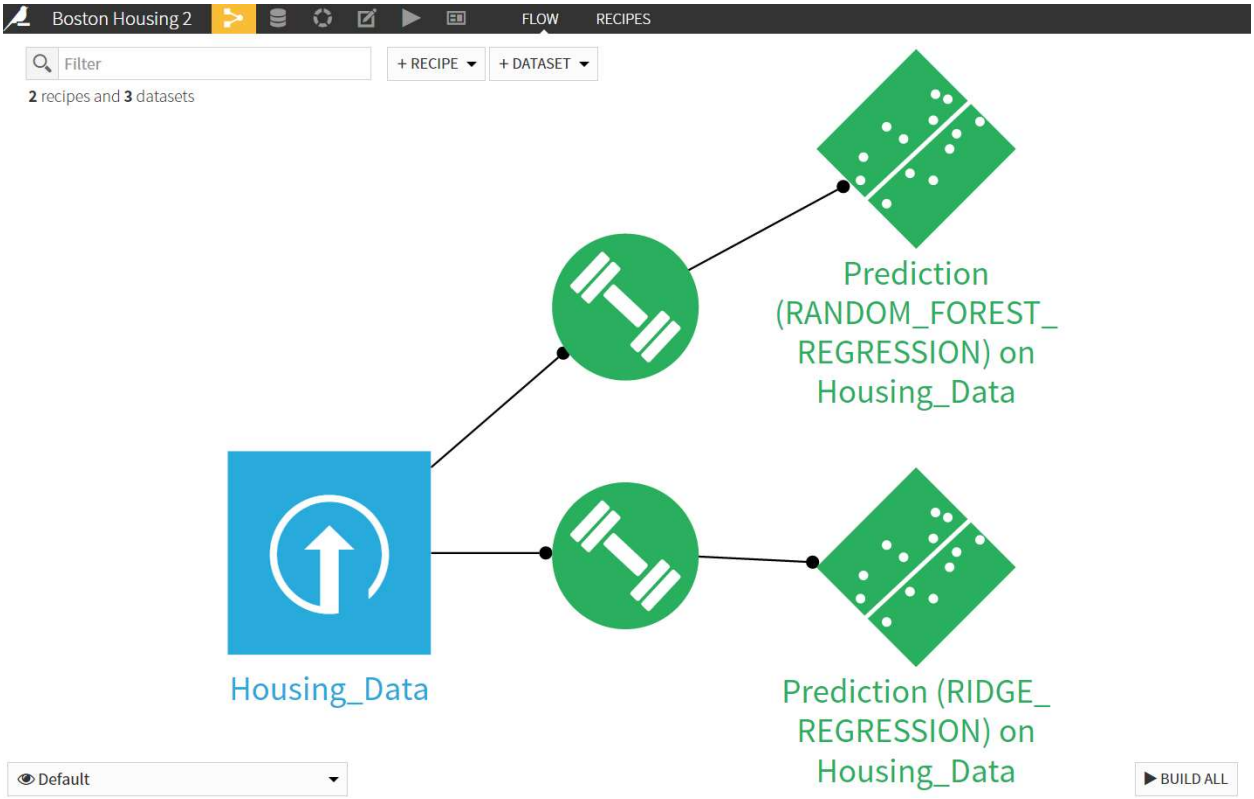


Detailed metrics of Ridge regression model:

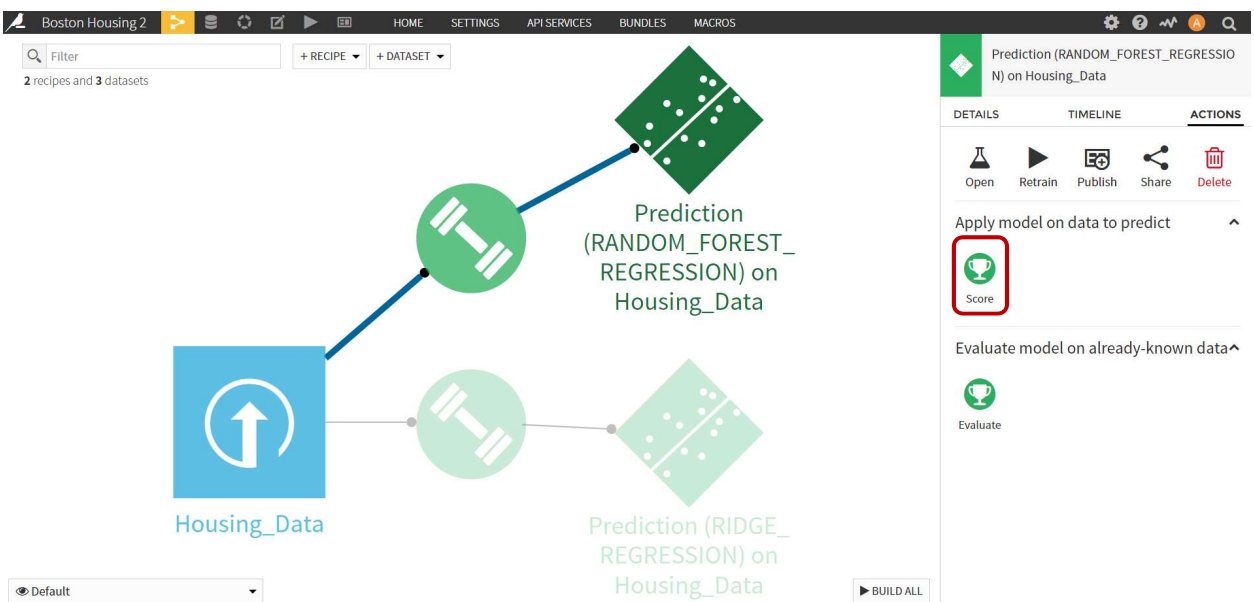


Considering all standard statistical scores, random forest models' predictions seem to be better than ridge regression model.

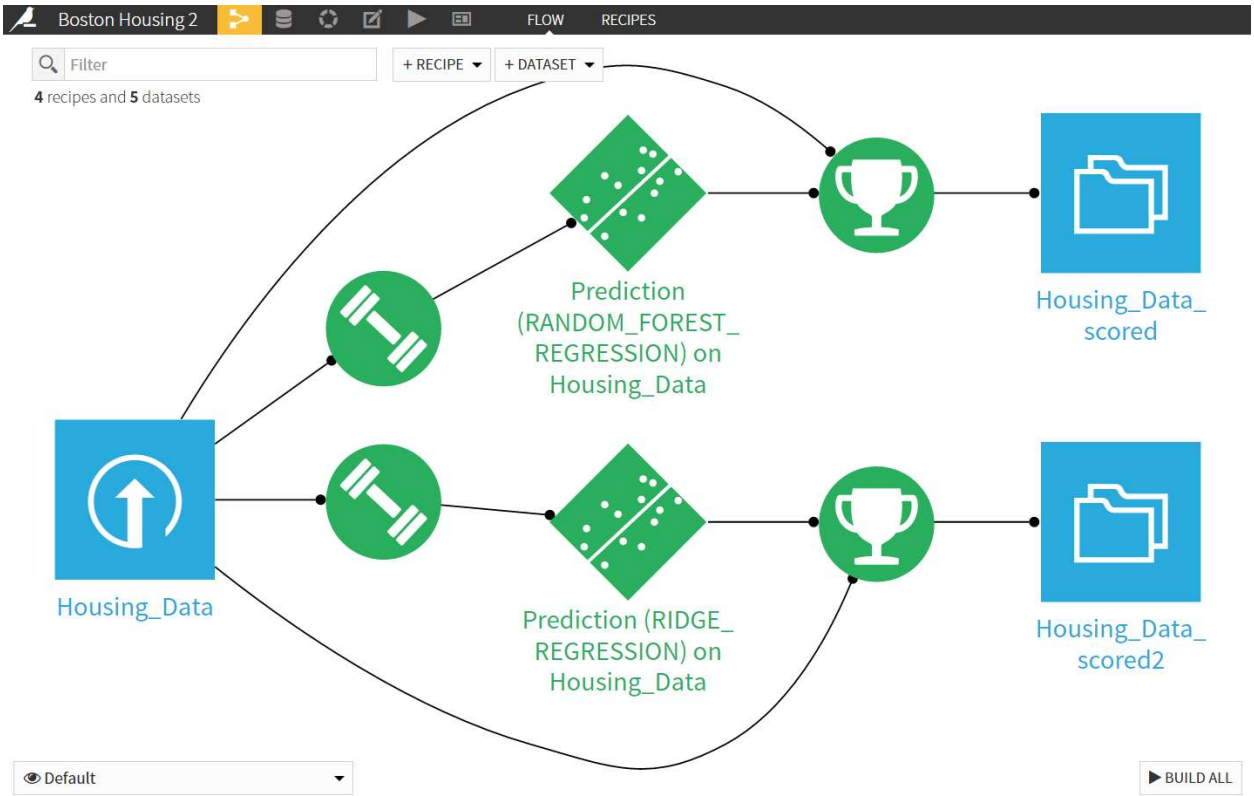
After deploying 2 models, the flow looks as shown below.



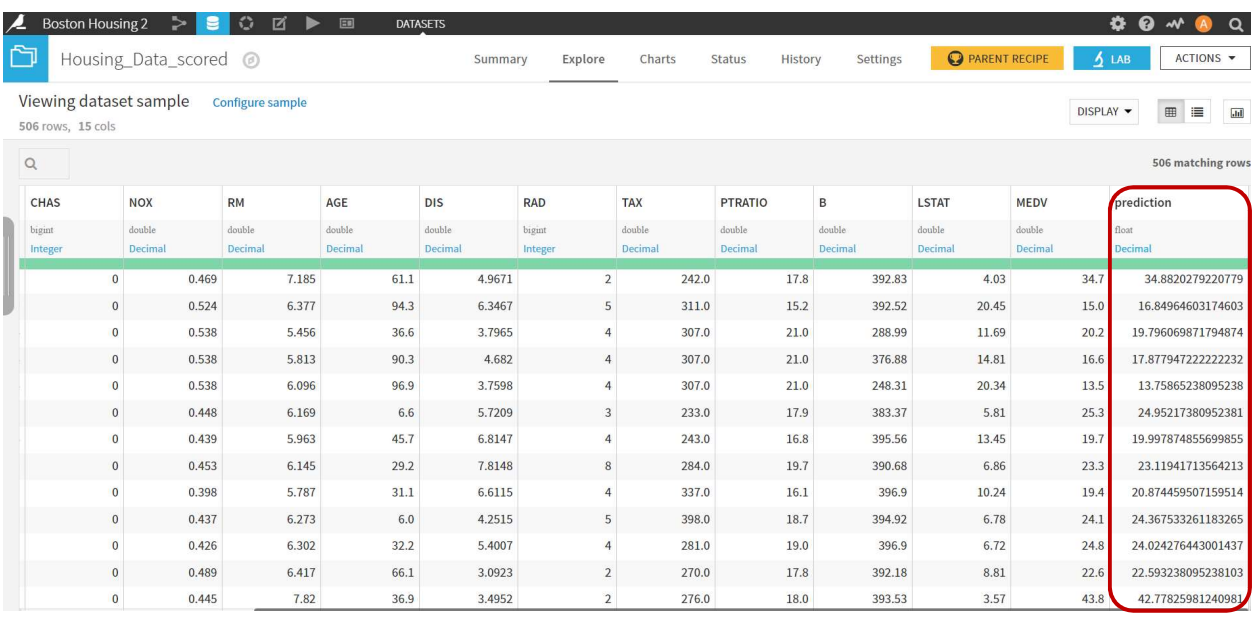
After clicking on a model, we can score it by using score recipe. This will apply the model on data to do predictions.



After applying this recipe to two models, the flow will look as shown below.



In scored datasets a new column called “prediction” appeared.



CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	prediction
bigint Integer	double Decimal	double Decimal	double Decimal	double Decimal	bigint Integer	double Decimal	double Decimal	double Decimal	double Decimal	double Decimal	float Decimal
0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7	34.8820279220779
0	0.524	6.377	94.3	6.3467	5	311.0	15.2	392.52	20.45	15.0	16.84964603174603
0	0.538	5.456	36.6	3.7965	4	307.0	21.0	288.99	11.69	20.2	19.796069871794874
0	0.538	5.813	90.3	4.682	4	307.0	21.0	376.88	14.81	16.6	17.877947222222232
0	0.538	6.096	96.9	3.7598	4	307.0	21.0	248.31	20.34	13.5	13.75865238095238
0	0.448	6.169	6.6	5.7209	3	233.0	17.9	383.37	5.81	25.3	24.95217380952381
0	0.439	5.963	45.7	6.8147	4	243.0	16.8	395.56	13.45	19.7	19.997874855699855
0	0.453	6.145	29.2	7.8148	8	284.0	19.7	390.68	6.86	23.3	23.11941713564213
0	0.398	5.787	31.1	6.6115	4	337.0	16.1	396.9	10.24	19.4	20.874459507159514
0	0.437	6.273	6.0	4.2515	5	398.0	18.7	394.92	6.78	24.1	24.367533261183265
0	0.426	6.302	32.2	5.4007	4	281.0	19.0	396.9	6.72	24.8	24.024276443001437
0	0.489	6.417	66.1	3.0923	2	270.0	17.8	392.18	8.81	22.6	22.593238095238103
0	0.445	7.82	36.9	3.4952	2	276.0	18.0	393.53	3.57	43.8	42.77825981240981

As my final submission dataset I choose the scored dataset prepared by using random forest model, since it is the best prediction model in the existing models according to the standard statistic metrics.

Final Submission Dataset (Other columns than MEDV and predictions are hided):

Housing_Data_scored
.csv

Conclusion

While being not sure about the accuracy of the results of my final project, I am very glad for having chance to learn many new things about the hot topic “Big Data” and learning how to use DSS.