

MACHINE LEARNING PROJECT REPORT - TEAM GEFORCE

NACHU, JAMIE, SHRUTHI, RIITU, VIJAYA

Problem Statement

- ❖ The Dataset given has historic data of houses sold between May 2014 to May 2015 at King County, Washington State, USA
- ❖ The training dataset consists of 21 attributes and 10000 rows. The testing dataset for which the price has to be predicted consists 20 rows.
- ❖ The aim is predict the sales of houses in King County with an accuracy of at least 85-90% (relative error less than 15%) and understand which factors are responsible for higher property value - \$650K and above

Targeted Solution and Our Role

Team Geforce is property agent team working for URA. Our goal is to predict the price of house as close to actual transaction prices as possible.

Some of the data provided such as zip code, year built, date of renovation, latitude and longitude are not understood by our algorithm. We did some feature engineering and created new attributes such as **Distance to City[#]**, **Area Score**, **House Age***, **Latest Date of Renovation** so that these new attributes can help our algorithm's learning better

* - House Age is calculated based on last renovated date.

- Distance to city is calculated assuming that capital city's latitude and longitude as 47, -122.



Modus Operandi:



DATA CLEANUP,
FEATURE
CONSTRUCTION
AND FEATURE
ENGINEERING

OPTIMIZE
SELECTION

OPTIMIZE
PARAMETERS

SELECT
ATTRIBUTES
AND
PARAMETER
TUNING

Data Cleanup and Feature Construction

- × Converted all the years and date columns to the date format by; setting it to Polynominal while importing the data and then changing it to the date format using the “Nominal to Date” operator.
- × Normalized the training set for the 6 “sgft” attributes
- × Engineered 4 new attributes using the generate attributes operator:

Distance to city

Area Score

Latest Date of Renovation

House Age



Views:

Design

Results

Turbo Prep

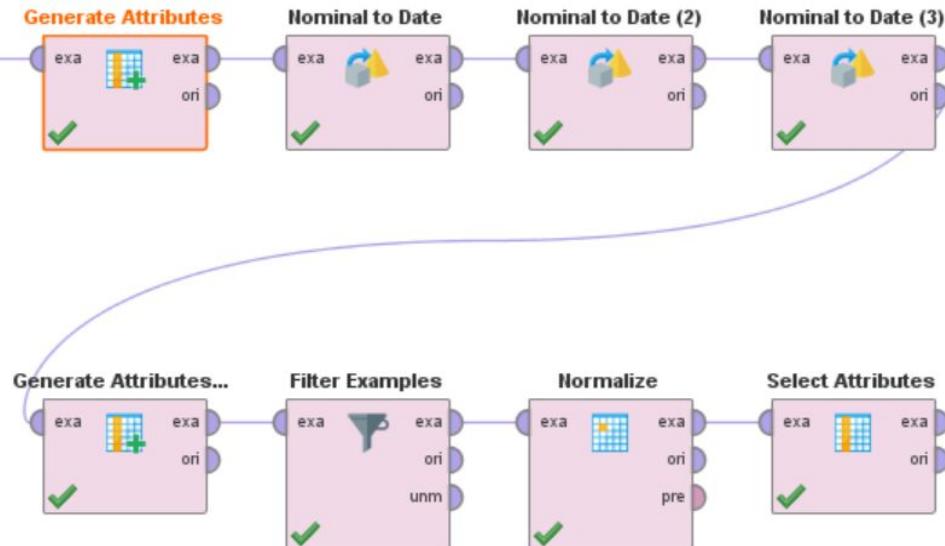
Auto Model

Find da

Process

1 Process ▶ Data Cleanup ▶

100%



Data Cleanup and Feature Construction

Feature Construction

Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
DistanceToCity	(47*-122)-(lat*long)
AreaScore	98199-zipcode

 Add Entry  Remove Entry  Apply  Cancel

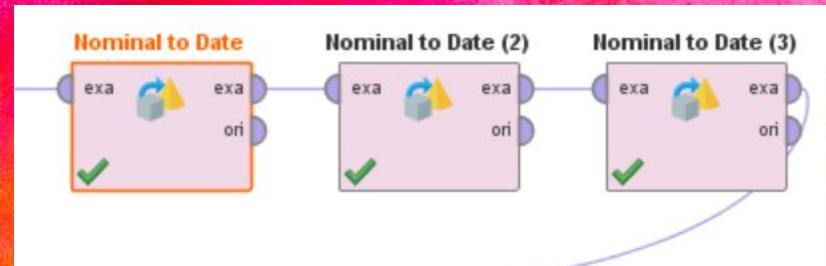
Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
HouseAge	date_diff(Renovated,date_modified)/1000/60/60/24/365

 Add Entry  Remove Entry  Apply  Cancel

Nominal to date



Parameters

Nominal to Date (2) (Nominal to Date)

attribute name	Renovated
date type	date
date format	yyy 12
time zone	SYSTEM
locale	English (United States...)

[Hide advanced parameters](#)

Parameters

Nominal to Date

attribute name	date_modified
date type	date
date format	yyyMMdd 12
time zone	SYSTEM
locale	English (United States...)

[Hide advanced parameters](#)

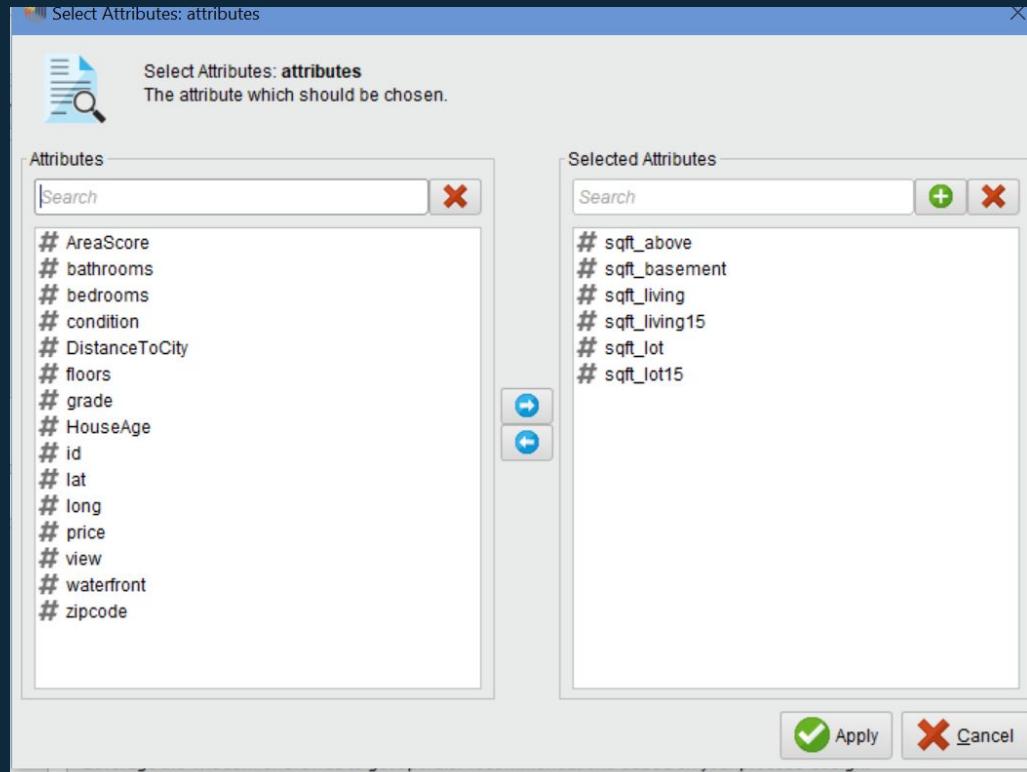
Parameters

Nominal to Date (3) (Nominal to Date)

attribute name	yr_built
date type	date
date format	yyy 12
time zone	SYSTEM
locale	English (United States...)

[Hide advanced parameters](#)

Normalizing only 6 attributes to maintain proportionality

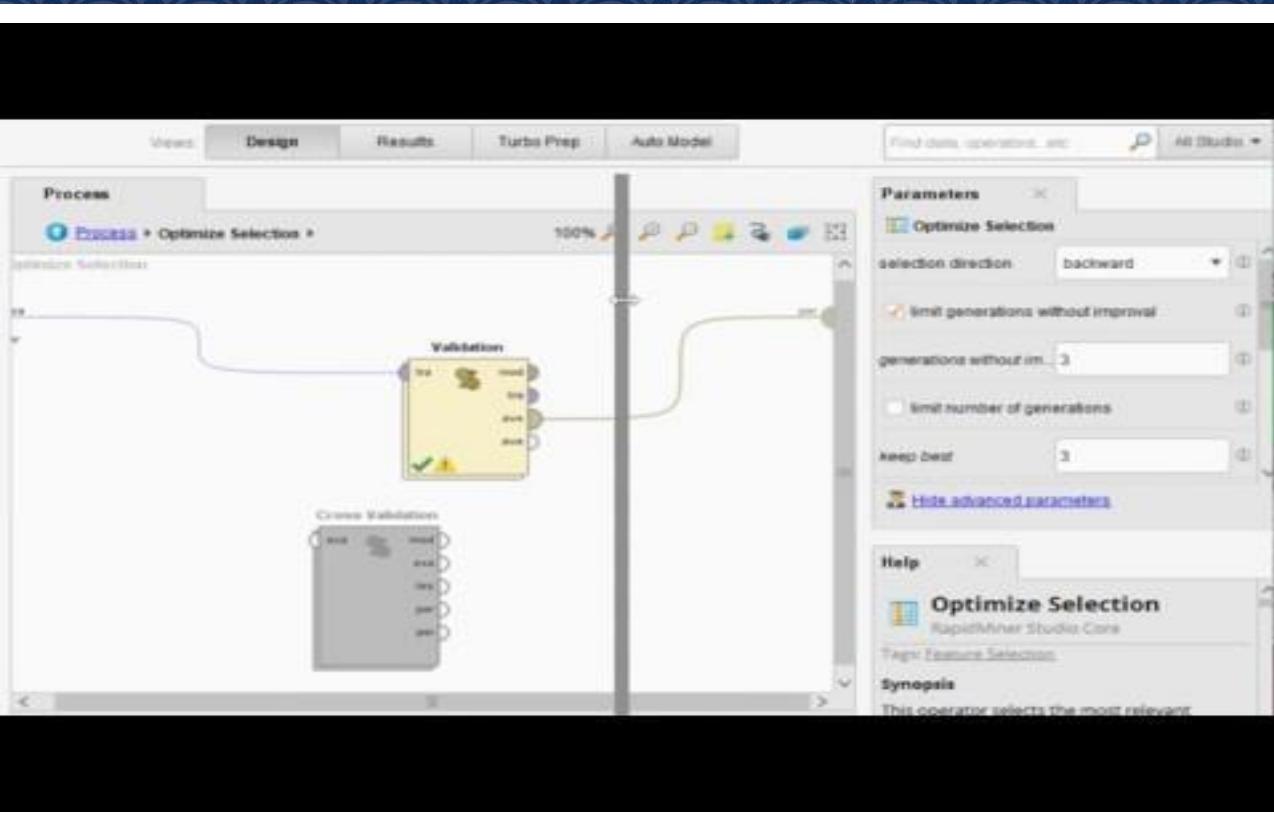


DATA CLEANUP RESULTS

	DistanceTo...	AreaScore	date_modifi...	Renovated	yr_built	HouseAge
	74.577	21	Oct 13, 2014	Jan 1, 1955	Jan 1, 1955	59.822
	103.185	74	Dec 9, 2014	Jan 1, 1991	Jan 1, 1951	23.953
	101.147	171	Feb 25, 2015	Jan 1, 1933	Jan 1, 1933	82.205
	82.213	63	Dec 9, 2014	Jan 1, 1965	Jan 1, 1965	49.970
	77.392	125	Feb 18, 2015	Jan 1, 1987	Jan 1, 1987	28.151
	80.282	146	May 12, 2014	Jan 1, 2001	Jan 1, 2001	13.367
	53.254	196	Jun 27, 2014	Jan 1, 1995	Jan 1, 1995	19.499
	64.893	1	Jan 15, 2015	Jan 1, 1963	Jan 1, 1963	52.074
	78.512	53	Apr 15, 2015	Jan 1, 1960	Jan 1, 1960	55.323
	46.413	161	Mar 12, 2015	Jan 1, 2003	Jan 1, 2003	12.200
	80.188	192	Apr 3, 2015	Jan 1, 1965	Jan 1, 1965	50.285
	98.105	84	May 27, 2014	Jan 1, 1942	Jan 1, 1942	72.449
	103.144	171	May 28, 2014	Jan 1, 1927	Jan 1, 1927	87.463



OPTIMIZE SELECTION



Done to choose the best combination of columns for our model.

RESULT:
*Relative error of
19.4% with 7
attributes*

Used Optimize Selection operator for choosing the best attribute combination to get optimal result

Result Individual Selection

All Individuals

Index	Features	Names	root_mean_squared...	relative_error ↑	squared_error	correlation
19	7	sqft_living, sqft_lot, waterfront, view, grade, DistanceToCity, AreaScore	182739.957	0.193	33393892040.783	0.892
2	7	sqft_living, bedrooms, floors, waterfront, grade, DistanceToCity, AreaScore	210563.330	0.194	44336916118.693	0.858
5	7	sqft_living, sqft_lot15, bedrooms, floors, grade, DistanceToCity, AreaScore	198411.955	0.198	39367303930.889	0.863
4	7	sqft_living, sqft_lot15, bedrooms, waterfront, grade, DistanceToCity, AreaScore	183967.109	0.199	33843897243.556	0.883
18	7	sqft_living, floors, waterfront, view, grade, DistanceToCity, AreaScore	177754.279	0.200	31596583759.155	0.892
10	7	sqft_living, floors, waterfront, view, grade, DistanceToCity, AreaScore	181162.519	0.208	32819858412.686	0.888
11	7	sqft_living, bedrooms, waterfront, view, grade, DistanceToCity, AreaScore	183954.310	0.209	33839188299.573	0.886
17	7	sqft_lot, floors, waterfront, view, grade, DistanceToCity, AreaScore	218301.090	0.212	47655365928.008	0.837
13	7	sqft_living, bedrooms, floors, waterfront, grade, DistanceToCity, AreaScore	184416.209	0.212	34009338214.935	0.883
22	7	sqft_living, sqft_lot, floors, waterfront, view, DistanceToCity, AreaScore	223912.255	0.212	50136697896.173	0.837
21	7	sqft_living, sqft_lot, floors, waterfront, grade, DistanceToCity, AreaScore	186201.187	0.213	34670881930.765	0.882
1	7	sqft_lot15, bedrooms, floors, waterfront, grade, DistanceToCity, AreaScore	228261.383	0.218	52103258973.142	0.816
20	7	sqft_living, sqft_lot, floors, view, grade, DistanceToCity, AreaScore	191030.472	0.219	36492641092.744	0.874
16	7	sqft_living, bedrooms, floors, waterfront, view, grade, DistanceToCity	189757.402	0.226	36007871780.455	0.866

Selected Individual

ExampleSet (1000 examples, 2 special attributes, 7 regular attributes)

Role	Name	Type	Statistics	Range	Missings
id	id	real	avg = 4392103485.002 +/- 28...	[9000025.000 ; 9842300485.0...]	0
label	price	integer	avg = 533477.350 +/- 411583....	[95000.000 ; 7700000.000]	0

Save Data... Select Cancel

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operat

Result History PerformanceVector (Performance) ExampleSet (Optimize Selection)

Open in Turbo Prep Auto Model Filter (1,000 / 1,000 examples): all

id	price	sqft_living	bedrooms	floors	waterfront	grade	DistanceTo...	AreaScore
6414100192	538000	0.188	3	2	0	7	103.185	74
2487200875	604000	0.135	4	1	0	7	82.213	63
7237550310	1230000	0.432	4	1	0	11	80.282	146
1321400060	257500	0.114	3	2	0	7	53.254	196
6054650070	400000	0.085	3	1	0	7	76.892	125
8091400200	252700	0.059	2	1	0	7	50.963	169
3303700376	667000	0.087	3	1.500	0	8	90.850	87
7895500070	240000	0.072	4	1	0	7	54.108	198
2078500320	605000	0.192	4	2	0	8	73.228	143
5547700270	625000	0.188	4	2	0	9	76.255	125
822039084	1350000	0.203	3	1	1	9	70.679	129
7231300125	345000	0.237	5	1	0	8	69.171	143
9822700295	885000	0.210	4	2	0	9	94.305	94

ExampleSet (1,000 examples, 2 special attributes, 7 regular attributes)

File Edit Process View Connections Settings Extensions Help

Views: Design Result History PerformanceVector (Performance) ExampleSet (Optimize Selection)

Performance

Criterion

- root mean squared error
- relative error
- squared error
- correlation

Description

relative_error

relative_error: 19.40% +/- 15.92%

File Edit Process View Connections Settings Extensions Help

Views: Design Result History PerformanceVector (Performance) ExampleSet (Optimize Selection)

Performance

Criterion

- root mean squared error
- relative error
- squared error
- correlation

Description

correlation

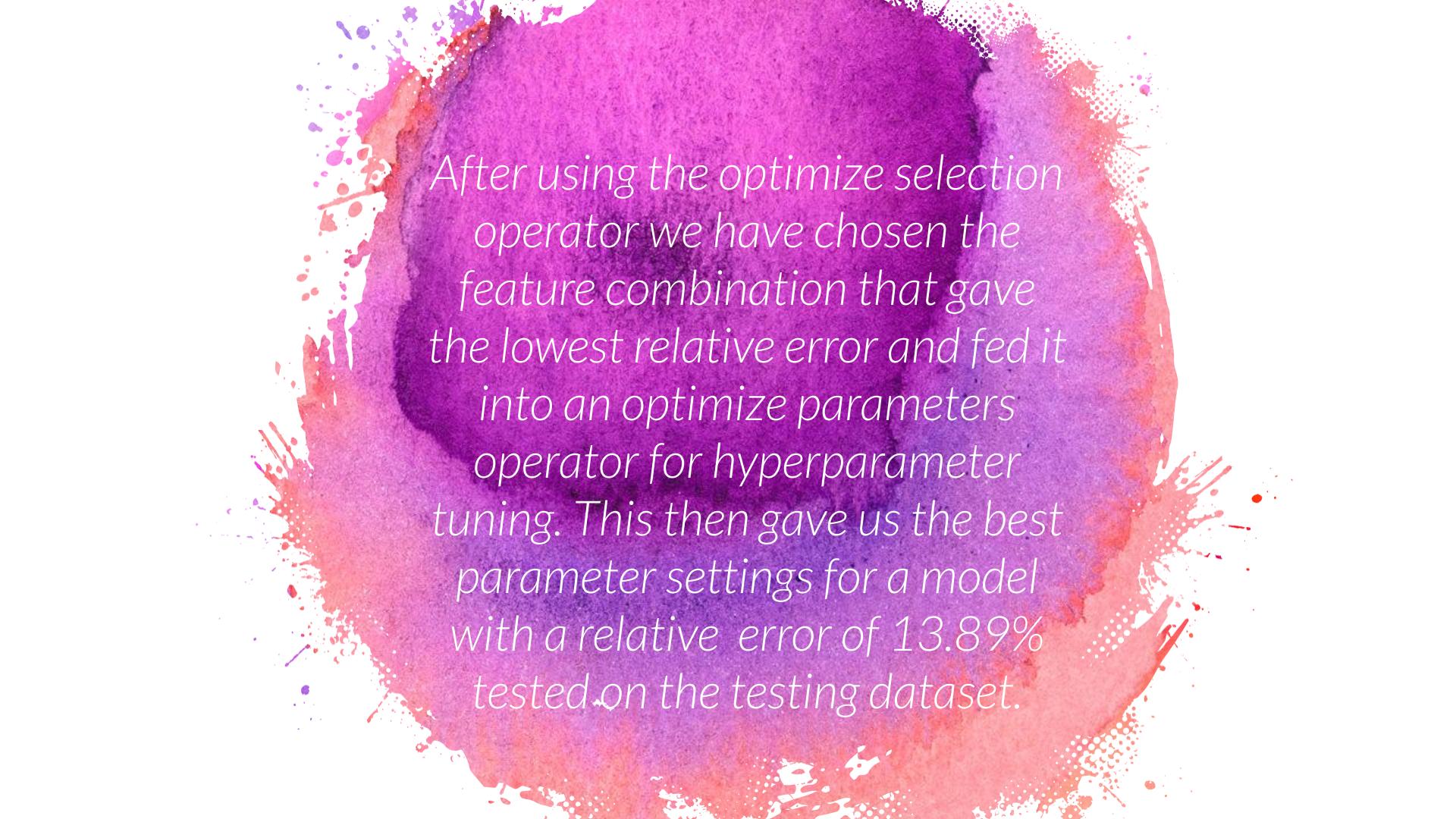
correlation: 0.858

Optimize Selection - Process Runthrough

The screenshot shows the RapidMiner interface with the 'Performance' tab selected. The main area displays a table titled 'ExampleSet (Optimize Selection)' containing 1,000 examples. The columns are: id, price, sqrt_listing, bedrooms, floors, waterfront, grade, DistanceTo..., and Arreactions. The first 10 rows are highlighted with a green background. To the left of the table, there are icons for Data, Models, Predictions, and Annotations. On the right, a 'Repository' panel lists various training resources. At the bottom, a status bar shows the search bar 'Type here to search:' and the date/time '12:19 PM 22-Apr-2015'.

ID	price	sqrt_listing	bedrooms	floors	waterfront	grade	DistanceTo...	Arreactions
64141001182	1000000	2.158	3	2	0	7	103.100	14
24870008779	4000000	0.156	6	1	0	7	82.213	0.3
72370002119	1200000	0.432	4	1	0	10	85.282	140
-13214000060	2000000	0.114	3	2	0	7	83.254	196
62540000073	4000000	0.089	5	1	0	7	76.862	125
60914000079	292700	0.058	2	1	0	7	80.863	149
33407000079	807000	0.067	3	1.000	0	8	90.850	87
78800000073	2400000	0.072	4	1	0	7	84.426	136
28780000022	4000000	0.102	6	2	0	8	73.226	167
15247000023	4200000	0.168	4	2	0	8	76.250	129
4223700004	1300000	0.203	2	1	0	7	70.879	129
7251000125	3400000	0.237	0	1	0	8	68.171	142
16227001285	4000000	0.219	4	2	0	8	84.300	94

ExampleSet (1,000 examples, 2 special attributes, 7 regular attributes)



After using the optimize selection operator we have chosen the feature combination that gave the lowest relative error and fed it into an optimize parameters operator for hyperparameter tuning. This then gave us the best parameter settings for a model with a relative error of 13.89% tested on the testing dataset.

File Edit Project View Connections Settings Preferences Help

Design Results Turbo Prep Auto Model

Import Data, Transform... All Models...

Project History ExamplesSet (Select Attributes)

View Model (Tree (X))

Optimize Parameters (Grid (Z))

Data Simple Charts Advanced Charts

Optimize Parameters (Grid (Z)) (288 rows, 7 columns)

ID	RecordID	Normalise (4) (method)	A_000 (0) (measures_top4)	A_000 (4) (scaled_measures)	root_mean_squared_error_0
24	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	92187.155
81	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	139992.034
81	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	139458.027
82	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	162446.067
100	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	195470.596
43	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	196343.003
91	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	196728.544
0	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	197105.824
82	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	198164.007
81	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	198738.722
82	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	199626.875
24	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	199475.766
24	proportion_transformation	MaxMeasures	None (0) (scaledDistance)	None (0) (scaledDistance)	199581.404

Repository Import Data

- Project - Confirmed Data Clean
- PROJECT - Optimal parameters
- PROJECT - TRAIN 1
- project perform - sum 20%
- PROJECT REFERENCE FINAL 3
- PROJECT SERVICE OPTIMAL
- project perform - clean
- project perform - clean IT SP
- project perform - clean 10%
- project perform - clean 10%
- project perform - train 10%
- project perform - predict
- project predictor
- Random Forest - West Boston
- split up cities
- Steering - model - sum 20%
- Soft Clustering - clean
- Train Deterministic
- train better prep

Type here to search: 12:06 PM 14 Apr 2015




Result History ExampleSet (Select Attributes) Vote Model (Vote (4))

PerformanceVector (Performance (3)) Optimize Parameters (Grid) (2)

Data

Optimize Parameters (Grid) (2) (288 rows, 7 columns)

iteration	k-NN (4).k	Neural ...	Normalize (4).method	k-NN (4).measure_types	k-NN (4).mixed_measure	root_mean_squared...
41	42	24	proportion transformation	MixedMeasures	MixedEuclideanDistance	92187.135
194	42	81	proportion transformation	NumericalMeasures	MixedEuclideanDistance	100062.024
193	3	81	proportion transformation	NumericalMeasures	MixedEuclideanDistance	100458.027
192	80	62	proportion transformation	NumericalMeasures	MixedEuclideanDistance	102446.667
52	3	100	proportion transformation	MixedMeasures	MixedEuclideanDistance	105475.508
45	80	43	proportion transformation	MixedMeasures	MixedEuclideanDistance	106343.003
51	80	81	proportion transformation	MixedMeasures	MixedEuclideanDistance	106728.544
39	80	5	proportion transformation	MixedMeasures	MixedEuclideanDistance	107105.624

Simple Charts

Advanced Charts

The 41st iteration with a k of 42 for the k-NN algorithm , 24 training cycles for the neural network, using the proportion transformation with the mixed euclidean distance in the mixed measures to give us our chosen root mean squared error of 92187.135

OUR FINAL TRAINED MODEL WITHOUT PREDICTIONS

S PerformanceVector (Performance (3)) X Optimize Parameters (Grid) (2) X

History ExampleSet (Select Attributes) X Vote Model (Vote (4)) X

Open in Turbo Prep Auto Model Filter (9,993 / 9,993 examples): all ▾

	id	price	sqft_living	sqft_lot	waterfront	view	grade	DistanceTo...	AreaScore
7129300520	221900	0.069	0.003	0	0	7	74.577	21	
6414100192	538000	0.188	0.004	0	0	7	103.185	74	
5631500400	180000	0.033	0.006	0	0	6	101.147	171	
2487200875	604000	0.135	0.003	0	0	7	82.213	63	
1954400510	510000	0.111	0.005	0	0	8	77.392	125	
7237550310	1230000	0.432	0.061	0	0	11	80.282	146	
1321400060	257500	0.114	0.004	0	0	7	53.254	196	
2008000270	291850	0.058	0.006	0	0	7	64.893	1	
2414600126	229500	0.120	0.004	0	0	7	78.512	53	
3793500160	323000	0.129	0.004	0	0	7	46.413	161	
1736800520	662500	0.272	0.006	0	0	8	80.188	192	
9212900260	468000	0.067	0.003	0	0	7	98.105	84	

< ExampleSet (9,993 examples, 2 special attributes, 7 regular attributes) >

Our final
relative error:
 $13.89\% \pm 10.90\%$

relative_error

relative_error: $13.89\% \pm 10.90\%$



Solution Designed

- ✓ Team Geforce of URA created a Machine Learning model using Split Validation and Voting on 4 algorithms such as k-NN, Linear Regression, Decision Tree and Neural Network to arrive at this solution
- ✓ With the sample data of 1000 rows we achieved 14.53% relative error and with the full training dataset (10000 rows) we achieved 13.89%
- ✓ Also, we achieved a Root Mean Squared Error of 92187.135
- ✓ With this combination of algorithms, parameter settings, generate attributes and attribute selections we are confident that Cross Validation will be able to achieve much better results than this in a high configuration server/laptop
- ✓ This model will be able to predict results as close as 85% to 90% accuracy for the expected price range of the houses in King County, Washington state, USA



Solution Designed

- Based on the analysis done by Team Geforce the factors responsible for higher property value - \$650K and above are:
 1. Sqft Living
 2. Sqft Lot
 3. Waterfront
 4. View
 5. Grade
 6. Distance to City
 7. Area Score