

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

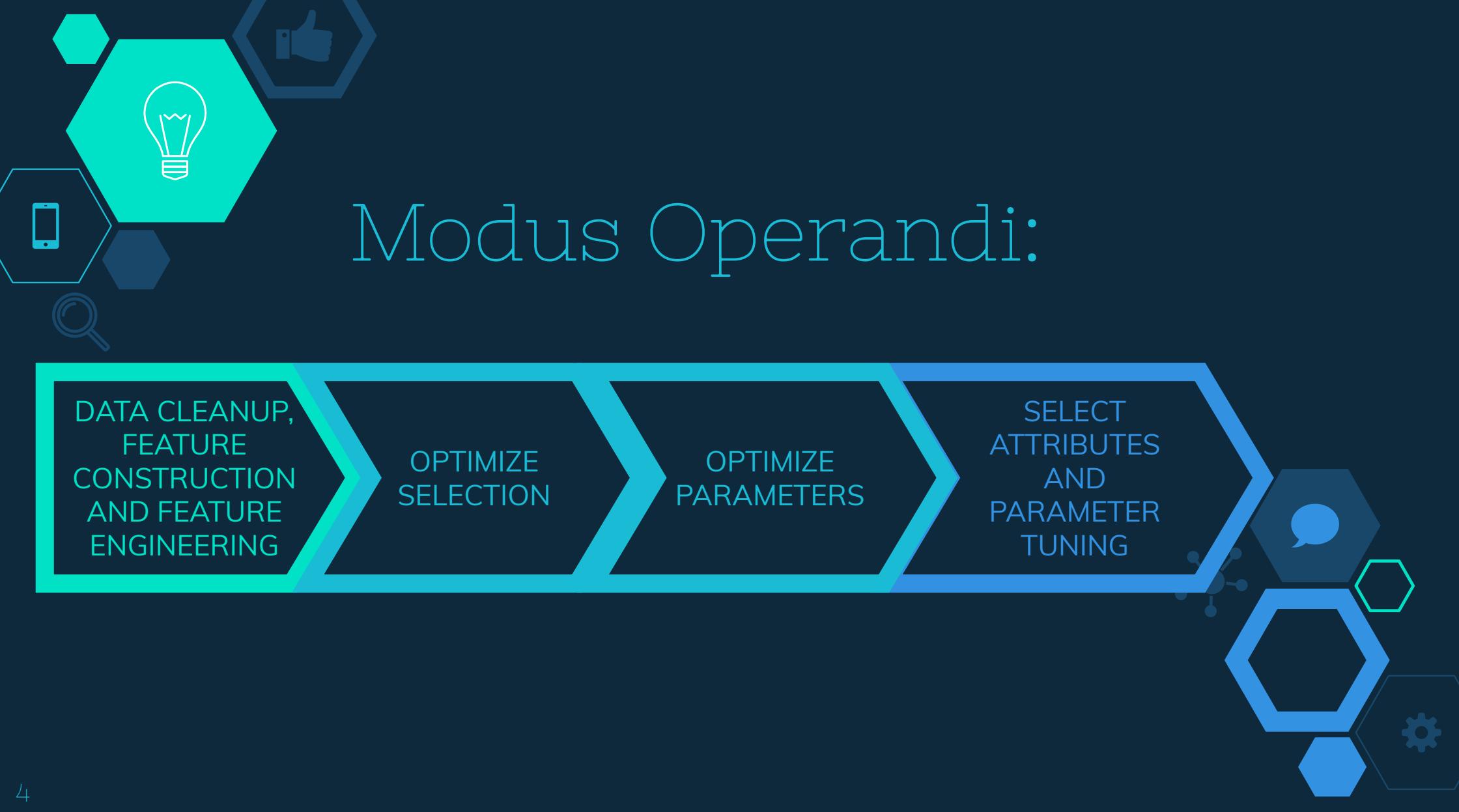
Targeted Solution and Our Role

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.



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

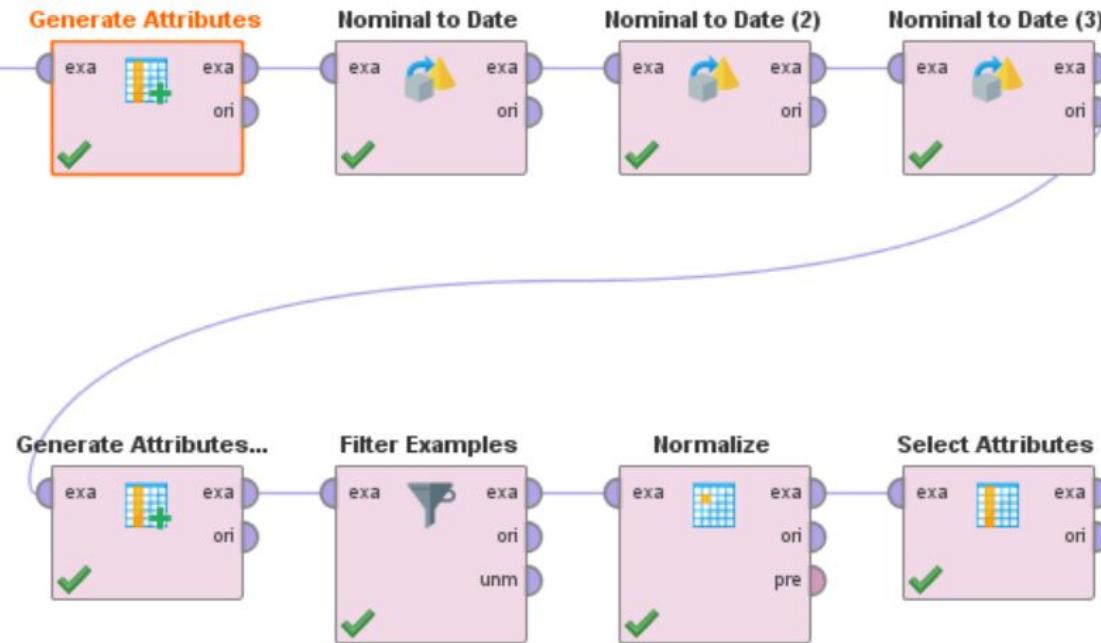
Auto Model

Find da

Process

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 | yyyy |
| time zone | SYSTEM |
| locale | English (United States) |

[Hide advanced parameters](#)

Parameters

Nominal to Date

| | |
|----------------|-------------------------|
| attribute name | date_modified |
| date type | date |
| date format | yyyyMMdd |
| time zone | SYSTEM |
| locale | English (United States) |

[Hide advanced parameters](#)

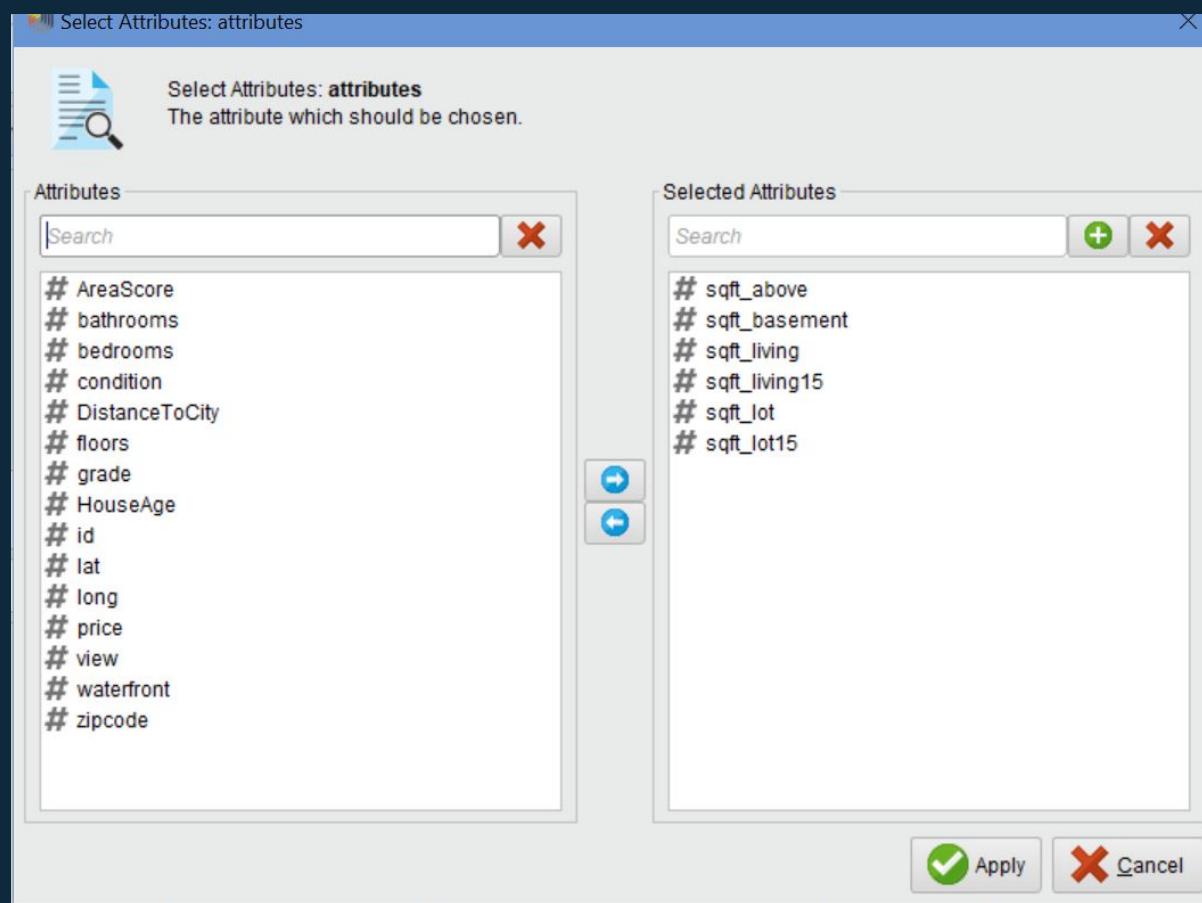
Parameters

Nominal to Date (3) (Nominal to Date)

| | |
|----------------|-------------------------|
| attribute name | yr_built |
| date type | date |
| date format | yyyy |
| time zone | SYSTEM |
| locale | English (United States) |

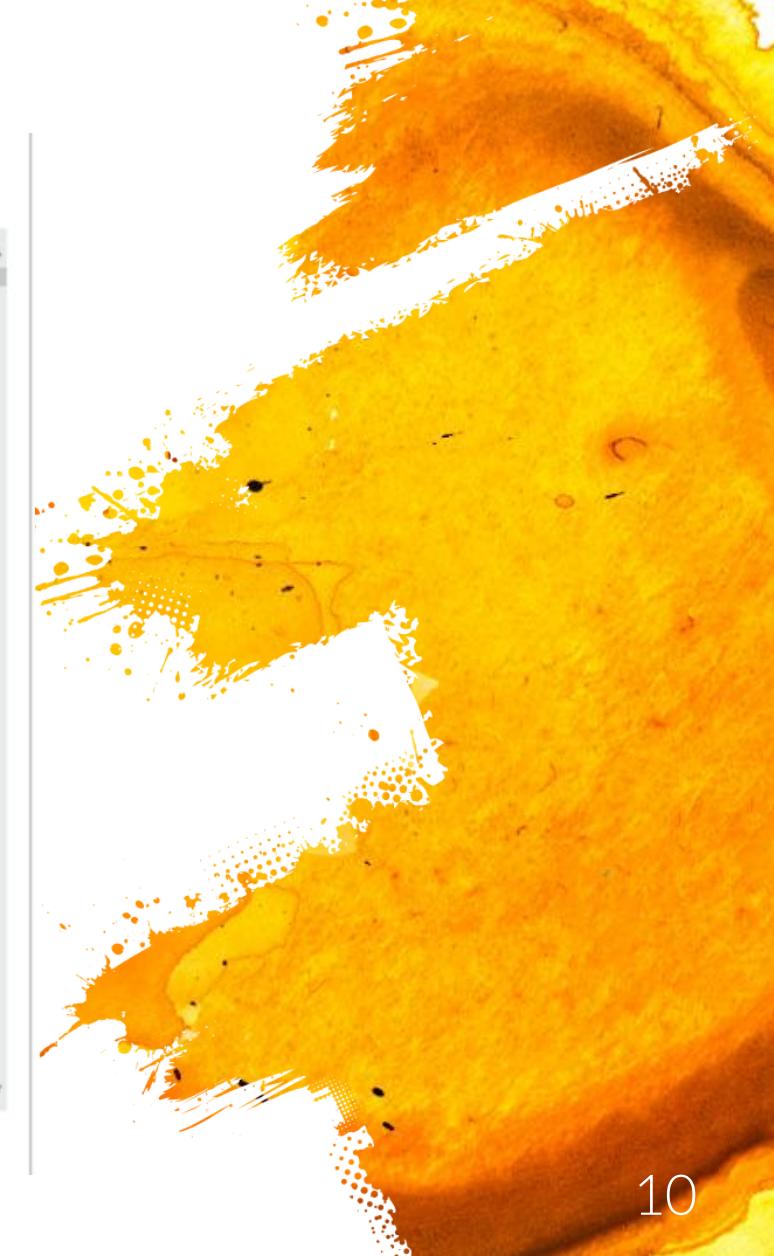
[Hide advanced parameters](#)

Normalizing only 6 attributes to maintain proportionality

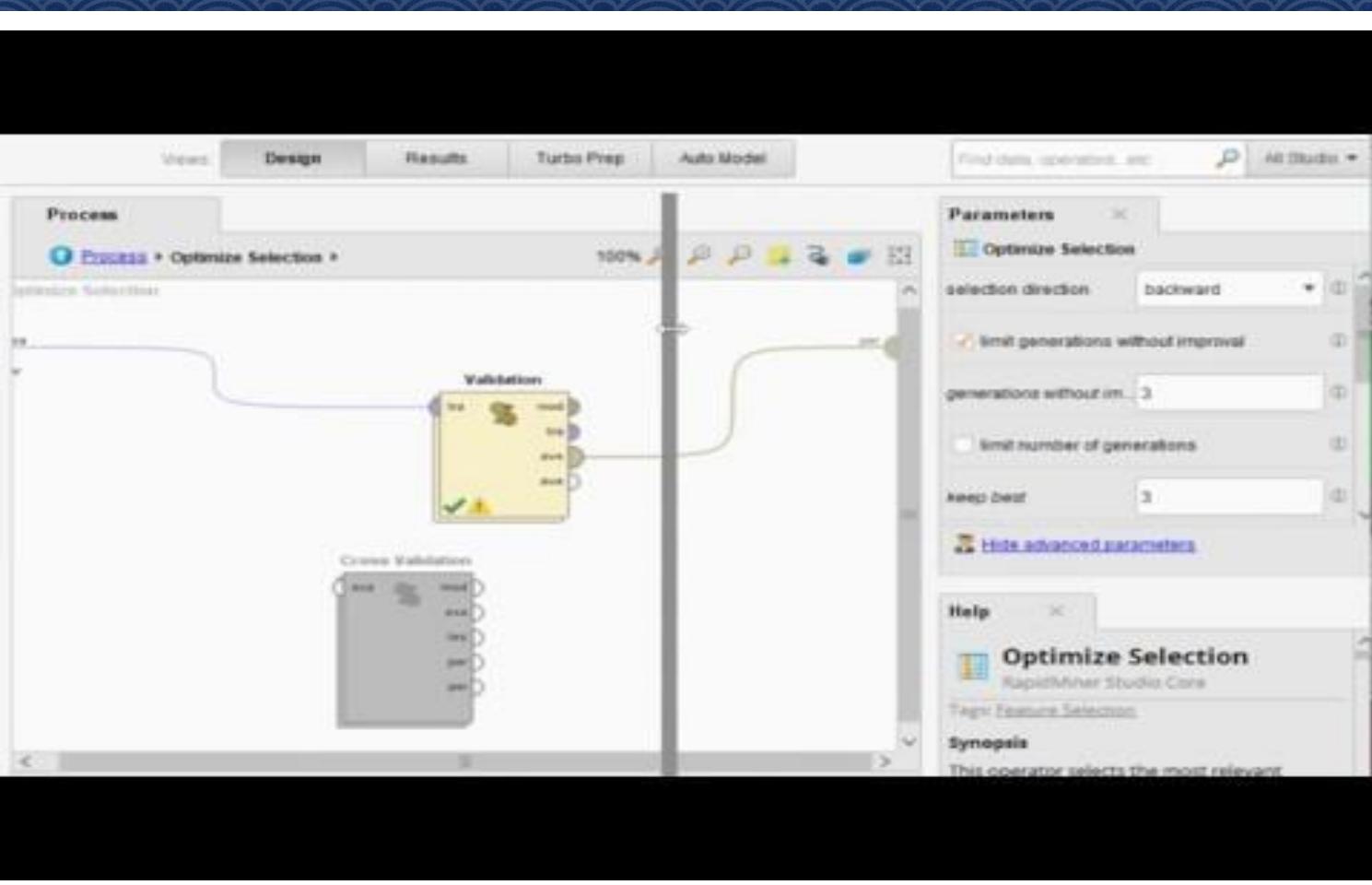


DATA CLEANUP RESULTS

| 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 | | | root_mean_squared... | relative_error ↑ | squared_error | correlation |
|-----------------|----------|---|----------------------|------------------|-----------------|-------------|
| Index | Features | Names | | | | |
| 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

//Local Repository/processes/PROJECT GEFORCE OPTIMIZE SELECTION FINAL – RapidMiner Studio Educational 9.3.000 @ SHRUTHI

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)

//Local Repository/processes/PROJECT GEFORCE OPTIMIZE SELECTION FINAL – RapidMiner Studio Educational 9.3.000 @ SHRUTHI

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)

Performance relative_error

Criterion root mean squared error relative error squared error correlation

relative_error: 19.40% +/- 15.92%

//Local Repository/processes/PROJECT GEFORCE OPTIMIZE SELECTION FINAL* – RapidMiner Studio Educational 9.3.000 @ SHRUTHI

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)

Performance correlation

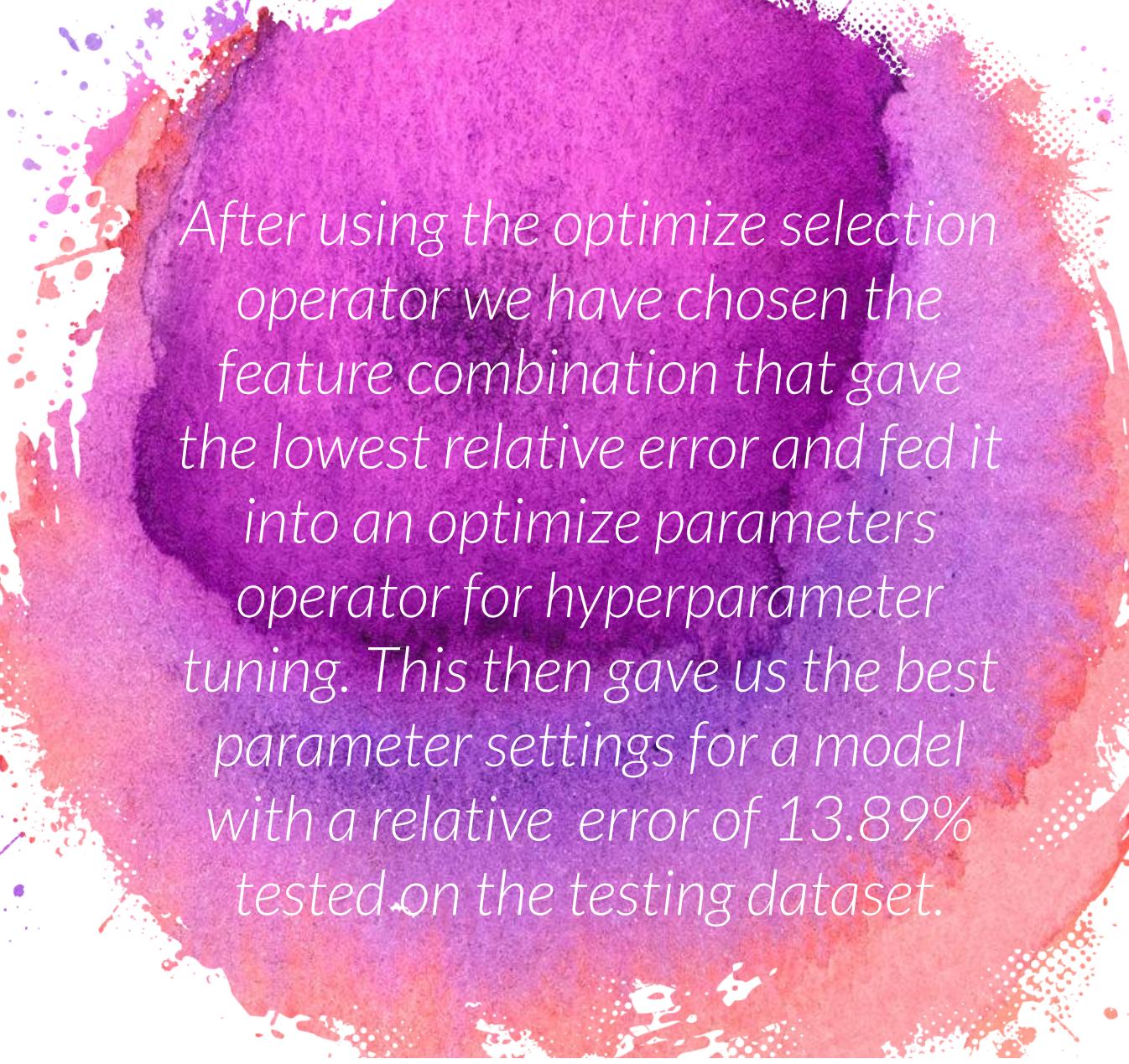
Criterion root mean squared error relative error squared error correlation

correlation: 0.858

Optimize Selection - Process Runthrough

The screenshot shows the RapidMiner interface with the 'ExampleSet (Optimize Selection)' view selected. The interface includes a toolbar at the top with various icons for file operations, a 'Design' tab, a 'Results' tab, a 'Turbo Prep' tab, and an 'Auto Model' tab. On the left, there's a sidebar with 'Result History', 'Data', 'Statistics', 'Visualizations', and 'Annotations' sections. The main area displays a table titled 'ExampleSet (Optimize Selection)' with columns: id, engine, wght_odor, instantane, Month, waterflow, grade, StatuscodeTc., and Anomalies. A green highlight covers the first 10 rows of the table. Below the table, a message says 'Exemplarset (1.000 examples, 12 special attributes, 17 regular attributes)'. The bottom of the screen shows the Windows taskbar with the Start button, a search bar, and pinned application icons.

| ID | engine | wght_odor | instantane | Month | waterflow | grade | StatuscodeTc. | Anomalies |
|--------------|-----------|-----------|------------|-------|-----------|-------|---------------|-----------|
| 54741001182 | 0.880000 | 0.1000 | 0 | 2 | 0 | 5 | 021.1000 | 54 |
| 24870000175 | 0.840000 | 0.1300 | 0 | 1 | 0 | 7 | 02.2110 | 63 |
| 72217000119 | 0.2300000 | 0.4300 | 0 | 1 | 0 | 11 | 00.2000 | 140 |
| 13214000000 | 20070000 | 0.1100 | 0 | 2 | 0 | 7 | 02.2504 | 136 |
| 40540000019 | 40000000 | 0.0800 | 0 | 1 | 0 | 7 | 00.0002 | 125 |
| 80914000000 | 20210000 | 0.0900 | 0 | 1 | 0 | 9 | 00.0003 | 140 |
| 34607000079 | 00700000 | 0.0600 | 0 | 13000 | 0 | 6 | 00.0001 | 87 |
| 78800000079 | 24000000 | 0.0372 | 0 | 1 | 0 | 7 | 04.1000 | 138 |
| 20780000000 | 00000000 | 0.1100 | 0 | 2 | 0 | 8 | 02.2200 | 143 |
| 155477000279 | 42000000 | 0.1800 | 0 | 2 | 0 | 6 | 00.2500 | 129 |
| 16220100044 | 13000000 | 0.2000 | 0 | 1 | 0 | 9 | 00.0079 | 126 |
| 12513001125 | 34000000 | 0.2200 | 0 | 1 | 0 | 8 | 00.1771 | 942 |
| 54027000295 | 00000000 | 0.2100 | 0 | 2 | 0 | 8 | 04.3000 | 84 |



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 Options View Connections Settings Extensions Help

Design Results Turbo Prep Auto Model

Result History

PerformanceVector (Performance [2])

Optimize Parameters (Grid [2])

Repository Import Data

Optimize Parameters (Grid [2])

| Row | Method | Measure (Measure) | Value (Selected Measure) | Value (Mean Absolute Error) |
|-----|---------------------------|-------------------|--------------------------|-----------------------------|
| 24 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 0.0187135 |
| 91 | proportion transformation | NumericalMeasure | Max(EuclideanDistance) | 1.004021234 |
| 91 | proportion transformation | NumericalMeasure | Max(EuclideanDistance) | 1.004561.027 |
| 92 | proportion transformation | NumericalMeasure | Max(EuclideanDistance) | 1.024461.007 |
| 100 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.054751.000 |
| 43 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.0623431.003 |
| 91 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.067231.044 |
| 91 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.071051.024 |
| 92 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.081041.007 |
| 91 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.097561.022 |
| 92 | proportion transformation | NumericalMeasure | Max(EuclideanDistance) | 1.106211.015 |
| 24 | proportion transformation | NumericalMeasure | Max(EuclideanDistance) | 1.154751.000 |
| 24 | proportion transformation | MaxMeasure | Max(EuclideanDistance) | 1.182811.000 |

Type here to search

12:14 PM 16 Apr 2019




The screenshot shows the KNIME interface with several tabs at the top: "Result History", "ExampleSet (Select Attributes)", "Vote Model (Vote (4))", "PerformanceVector (Performance (3))", and "Optimize Parameters (Grid) (2)". The "Optimize Parameters (Grid) (2)" tab is active, displaying a table titled "Optimize Parameters (Grid) (2) (288 rows, 7 columns)". The table has columns: iteration, k-NN (4).k, Neural ... (partially visible), Normalize (4).method, k-NN (4).measure_types, k-NN (4).mixed_measure, and root_mean_squared_. The first row is highlighted in blue, showing values: iteration 41, k-NN (4).k 42, Neural ... 24, Normalize (4).method proportion transformation, k-NN (4).measure_types MixedMeasures, k-NN (4).mixed_measure MixedEuclideanDistance, and root_mean_squared_ 92187.135.

| 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 |

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

| | 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% +/- 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