PREDICTING THE HOUSE PRICE IN KING COUNTY, USA

1. Project objective

The main goal of the project is to construct a predicting model for the house prices in King County, USA. Whether we are going to buy or sell a house, we need to estimate its price to make a better decision. And Orange data mining software can give us an accurate estimate.

2. Data understanding

The dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

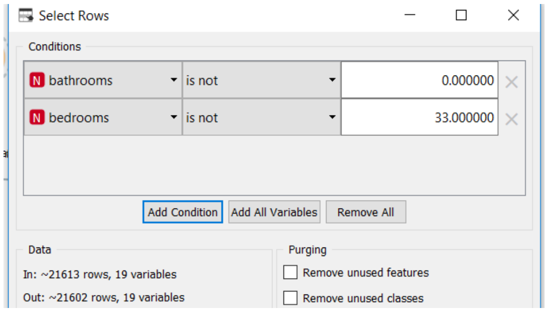
Data source: Kaggle (https://www.kaggle.com/harlfoxem/housesalesprediction/home)

Data description: 21600 records with 14 attributes

- id = a notation for a house
- **date** = Date house was sold (Categorical)
- price = Price is prediction target
- bedrooms = Number of Bedrooms/House
- bathrooms = Number of bathrooms
- sqft_living = square footage of the home
- sqft_lot = square footage of the lot
- floors = Total floors (levels) in house
- **waterfront** = House which has a view to a waterfront (categorical)
- view = Has been viewed
- condition = How good the condition is overall
- grade = overall grade given to the housing unit, based on King County grading system
- sqft_above = square footage of house apart from basement
- sqft_basement = square footage of the basement
- yr_built = Built Year
- yr_renovated = Year when house was renovated
- **zipcode** = zipcode (categorical)
- lat = Latitude coordinate
- long = Longitude coordinate
- sqft_living15 = Living room area in 2015(implies– some renovations) This might or might not have affected the lotsize area
- sqft_lot15 = lotSize area in 2015 (implies– some renovations)

3. Data Preparation

- Convert Yr_built and Yr_renovated to House_age and House_NewAge in Excel
- Statistics features: checking missing value, min, max. There is no missing value in the data set. But I noticed some weird values (33 bedrooms in one house and no bathrooms in some houses….)
- Select row: I eliminated the records with wrong values

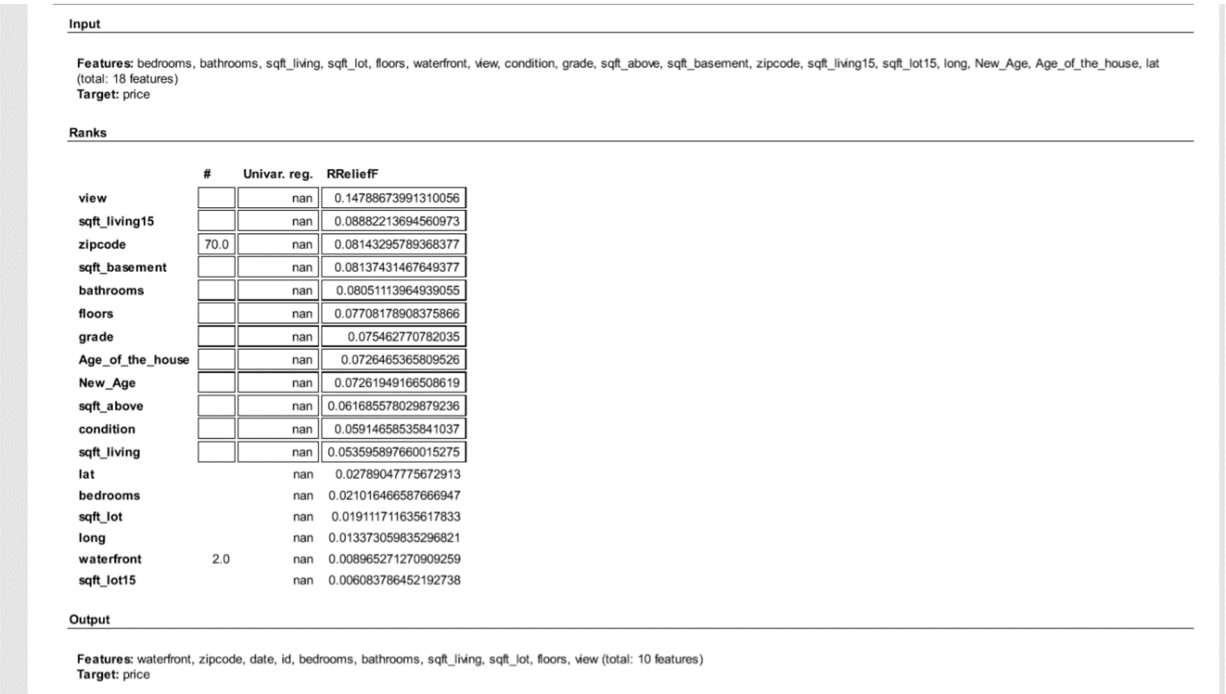


N bathrooms	is not	0.000000
N bedrooms	is not	33.000000

Data
In: ~21613 rows, 19 variables
Out: ~21602 rows, 19 variables

Purging
☐ Remove unused features
☐ Remove unused classes

- Select column: this function helped me set role for the target feature (price) and eliminate some unnecessary attibutes (Id and Date)
- Rank: using this widget to score the attributes according to their correlation with the price.


Input

Features: bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, zipcode, sqft_living15, sqft_lot15, long, New_Age, Age_of_the_house, lat (total: 18 features)
Target: price

Ranks

	#	Univar. reg.	RReliefF
view		nan	0.14788673991310056
sqft_living15		nan	0.08882213694560973
zipcode	70.0	nan	0.08143295789368377
sqft_basement		nan	0.08137431467649377
bathrooms		nan	0.08051113964939055
floors		nan	0.07708178908375866
grade		nan	0.075462770782035
Age_of_the_house		nan	0.0726465365809526
New_Age		nan	0.07261949166508619
sqft_above		nan	0.061685578029879236
condition		nan	0.05914658535841037
sqft_living		nan	0.053595897660015275
lat		nan	0.02789047775672913
bedrooms		nan	0.021016466587666947
sqft_lot		nan	0.019111711635617833
long		nan	0.013373059835296821
waterfront	2.0	nan	0.008965271270909259
sqft_lot15		nan	0.006083786452192738

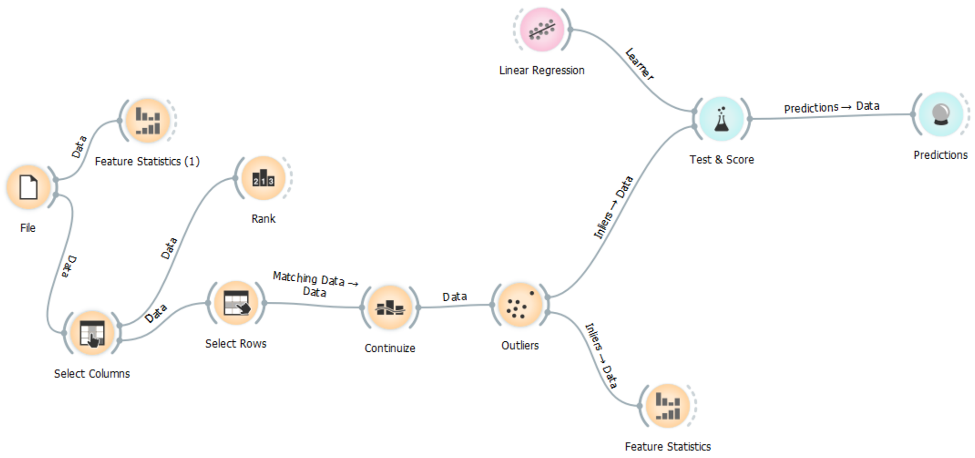
Output

Features: waterfront, zipcode, date, id, bedrooms, bathrooms, sqft_living, sqft_lot, floors, view (total: 10 features)
Target: price

- Continuze: to convert the categorical variables to dummies variables. In this case, the water front has 2 dummies variables and the zipcode has 70 variables.
- Outliers: helped me delete 217 outliers. Now, the data set has 21385 instances

4. Modeling

As the goal of this project is to predict how much a house should be sold, I chose Linear regression because it outputs the value of the variable as its prediction. The logistics regression can not be used in this case as it only outputs the probability of occurrance of an event as its prediction, for example, it can answer either Yes or No when being questioned whether a house should be sold at certain amount.



5. Results

Model	Details	MSE	RMSE	MAE	R2
Linear Regression	with 12 attributes filtered from Rank	27974455439.306	27974455439.306	101265.260	0.788
Linear Regression	all selected features	25970596539.379	161153.953	96320.299	0.807
Ridge Regression	alpha = 1	25966489724.253	161141.211	96278.304	0.807
Ridge Regression	alpha = 100	29202276362.621	170886.735	101017.186	0.783
Ridge Regression	alpha = 1000	40696507192.143	201733.753	124411.262	0.697
Lasso Regression	alpha = 1	25965721164.449	161138.826	96272.732	0.807
Lasso Regression	alpha = 100	26067523286.259	161454.400	96179.150	0.806
Lasso Regression	alpha = 1000	30667578590.013	175121.611	107794.491	0.772

Note:

- Sampling: using cross validation with number of folds: 10
- Lasso regression minimizes a penalized version of the least squares loss function with L1-norm penalty and Ridge regularization with L2-norm penalty.
- Alpha is the regularization strength. Regularization is designed to address the problem of overfitting and undefitting. To start with the overfitting, it means high variance and it is usually caused by a complicated function that creates a lot of unnecessary curves and angles unrelated to the data. This function fits to training data well but might cause poor results for the test set. On the

other hand, underfitting means low variance and a very simple model. This might also cause poor results too.

6. Reflection

What went well

- The data is quite in good condition, no missing data, most of them are numberic
- Orange tutorials on YouTube, the Orange blog, and the Kaggel

What did not go Well:

- Statistics: As I took the statistics class 4-5 years ago and almost forgot anything
- The Zip Code: numberic/categorical value
- Pick a dataset: I tried so many datasets before coming up with the King County House Price Prediction. For example, the LA restaurants violations, USA Name Datasets, Credit Card Fraud Detection, CryptoCurrencyPrice …. Honestly I find it difficult to both learn about the Orange tools and the new dataset with unfamiliar business problem at the same time.