DALHOUSIE UNIVERSITY

Data Mining In

HOME PRICE PREDICTION

BY

ARMAN ANDISHEH

JAFAR PASHAMI

DATA MINING AND MACHINE LEARNING

Dr. AHMED BALFAGIH

CONTENT TABLE

- Abstract
- Research questions
- Related works
- Methodology
- Data gathering
- Dataset description
- Results and Models
- Models Evaluation
- Conclusion
- Future works

- Housing Crisis in Halifax
- Planning and budgeting for owning home
- Deal with Fair price
- Take advantage of data mining in price prediction
  - Redfin site traffic

ABSTRACT

- Predict a reliable home price through data mining techniques?

- Which attributes do affect the home price?

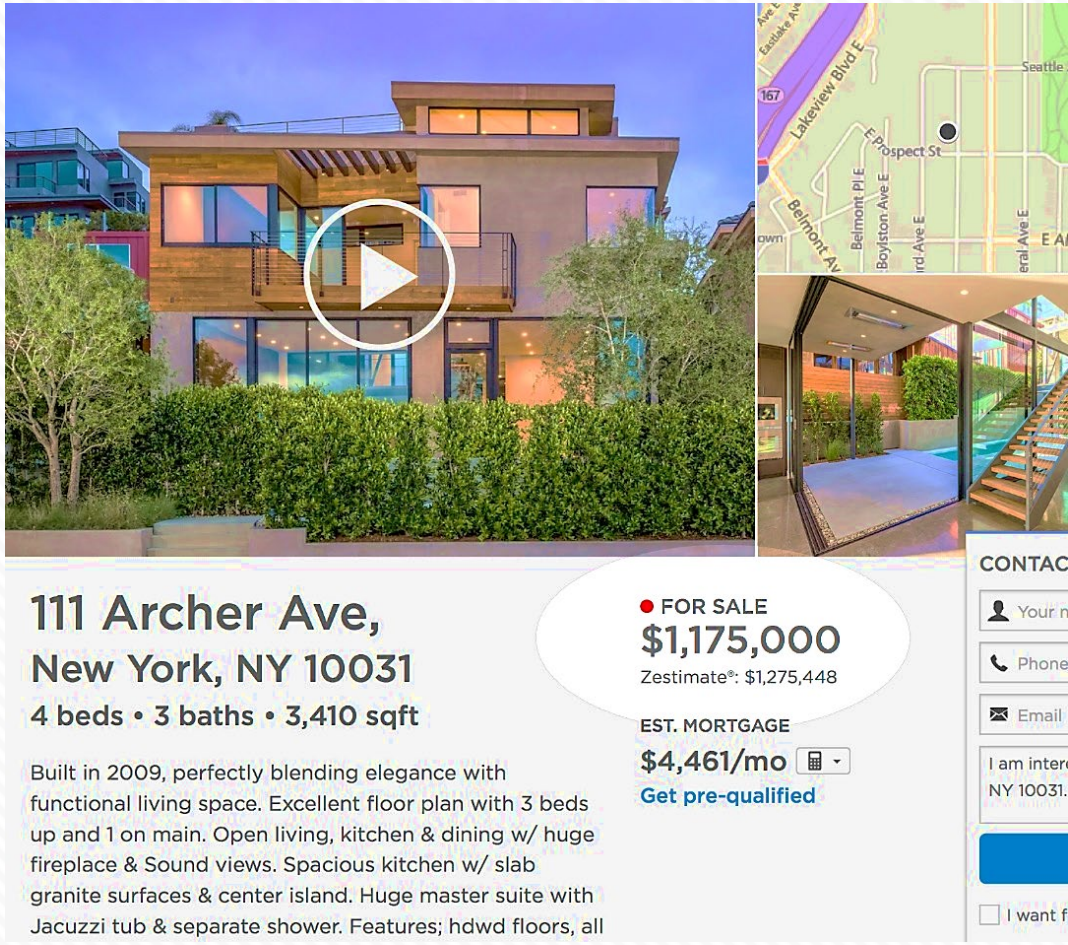
RESEARCH QUESTIONS

- Machine Learning techniques
  - Linear regression as best fit technique
  - Random forest and Decision tree
  - K-nearest

- Challenges
  - High missing values (K-nearest is reliable algorithm)
  - Attribute distribution
  - Time taking in data gathering
  - Combination of nominal and numerical values
  - Seasonal/ time error
  - Correlation between attributes make redundancy

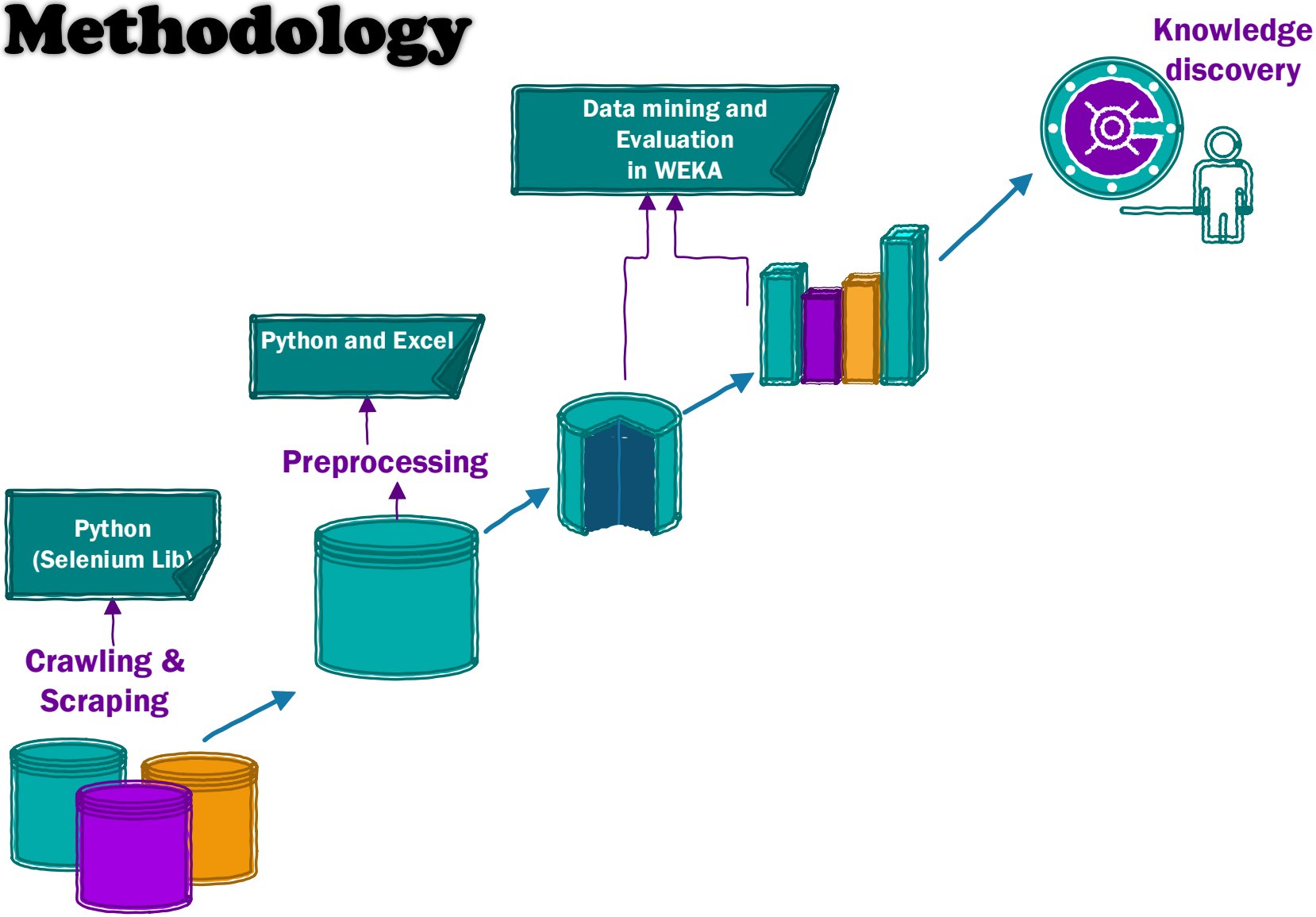
RELATED WORKS

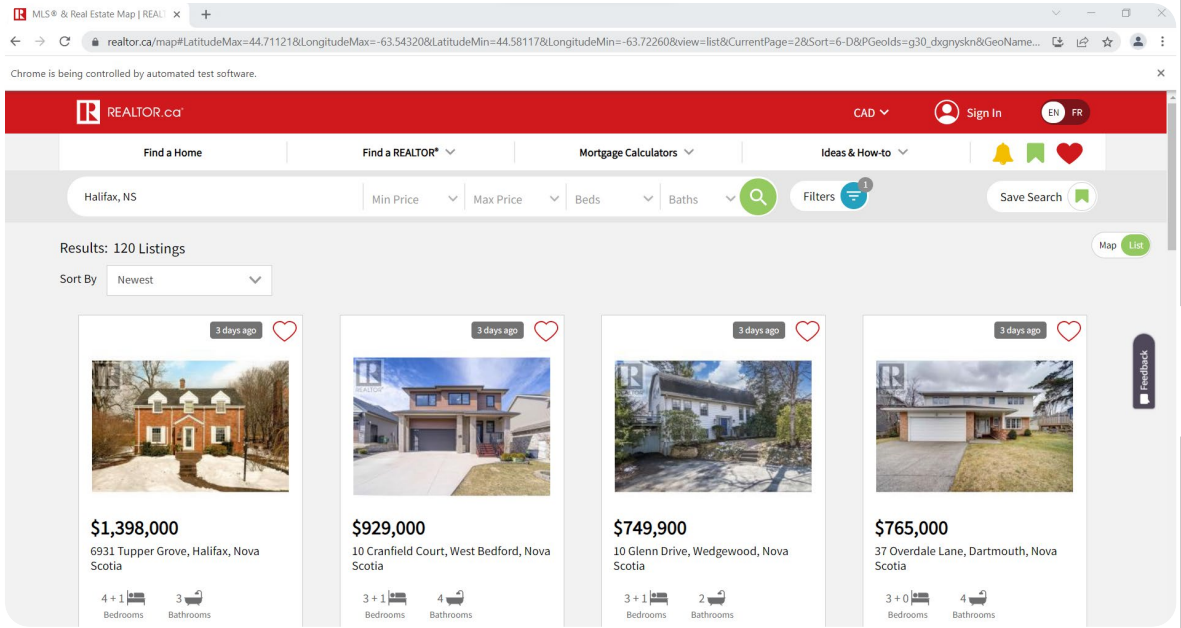
- **www.zillow.com**
  - Estimation error below 5%
  - Large data base, according Zillow is 7.5 m.
  - 58 Attributes
  - $ 1 million award for medizing log error
  - Decision tree and linear reg
- **www.redfin.com**
  - Random forest and gradient boosting.
  - Automating the Comparative Market Analysis (CMA)

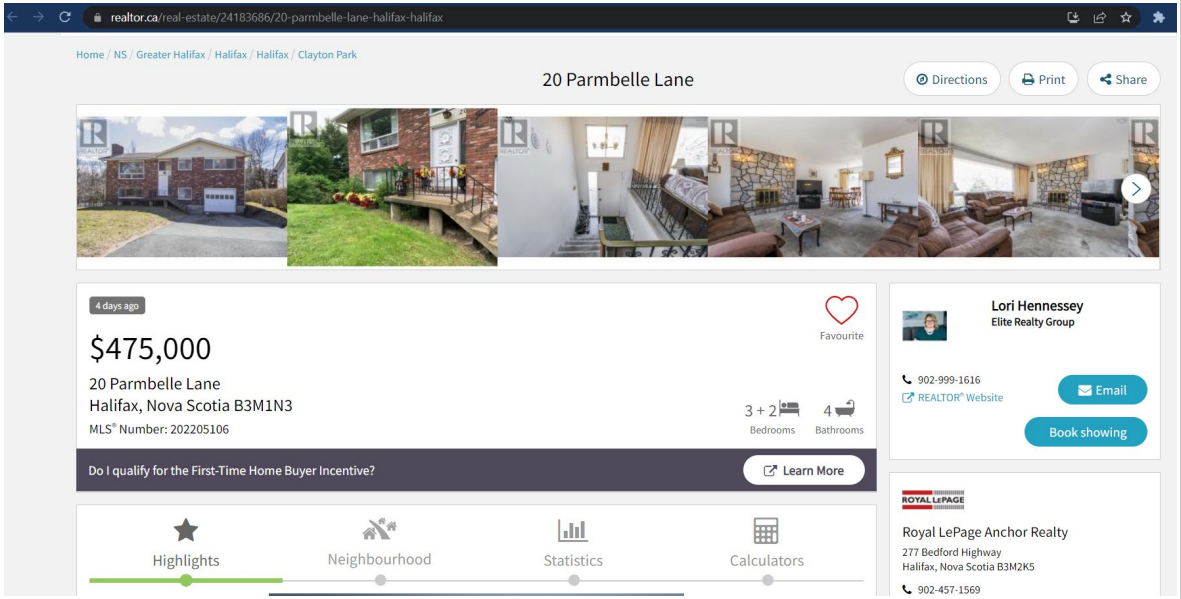


RELATED WORKS

Methodology







DATA GATHERING

- Owned by Canadian Real Estate Association (CREA)
- With over 240 million visits each year
- https://www.realtor.ca/

Tool for Scraping: Selenium library from python

First Collecting links

Collecting content of each page (28 columns)

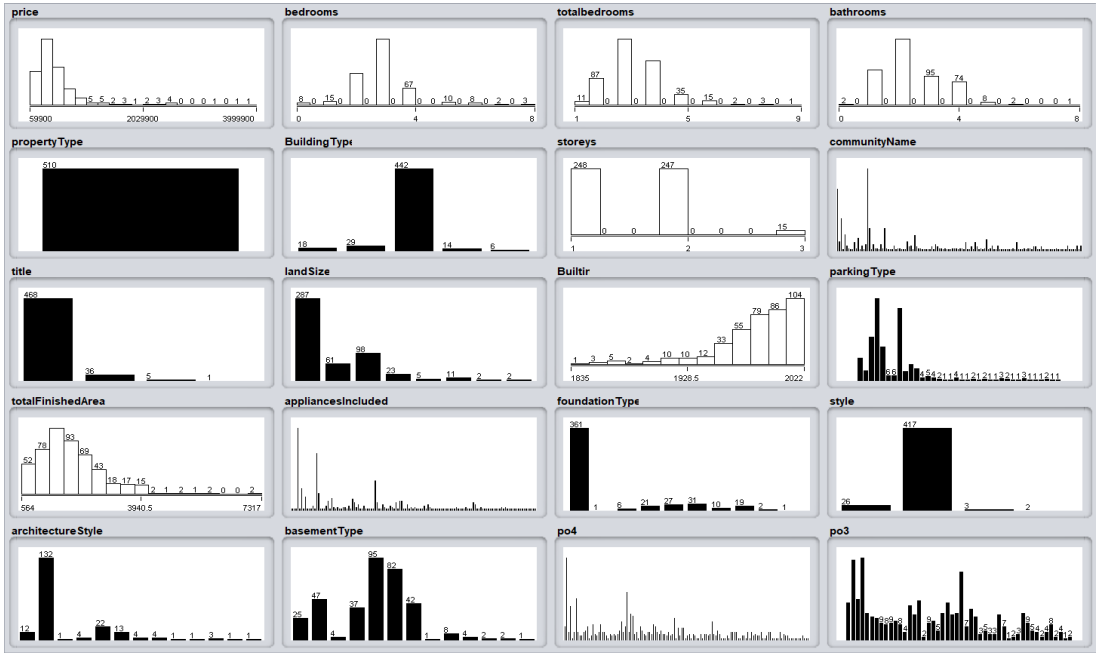
Saving results in CSV file

Cleaning CSV file using Pandas from python

Scraping and Preprocessing Data

DATASET DESCRIPTION

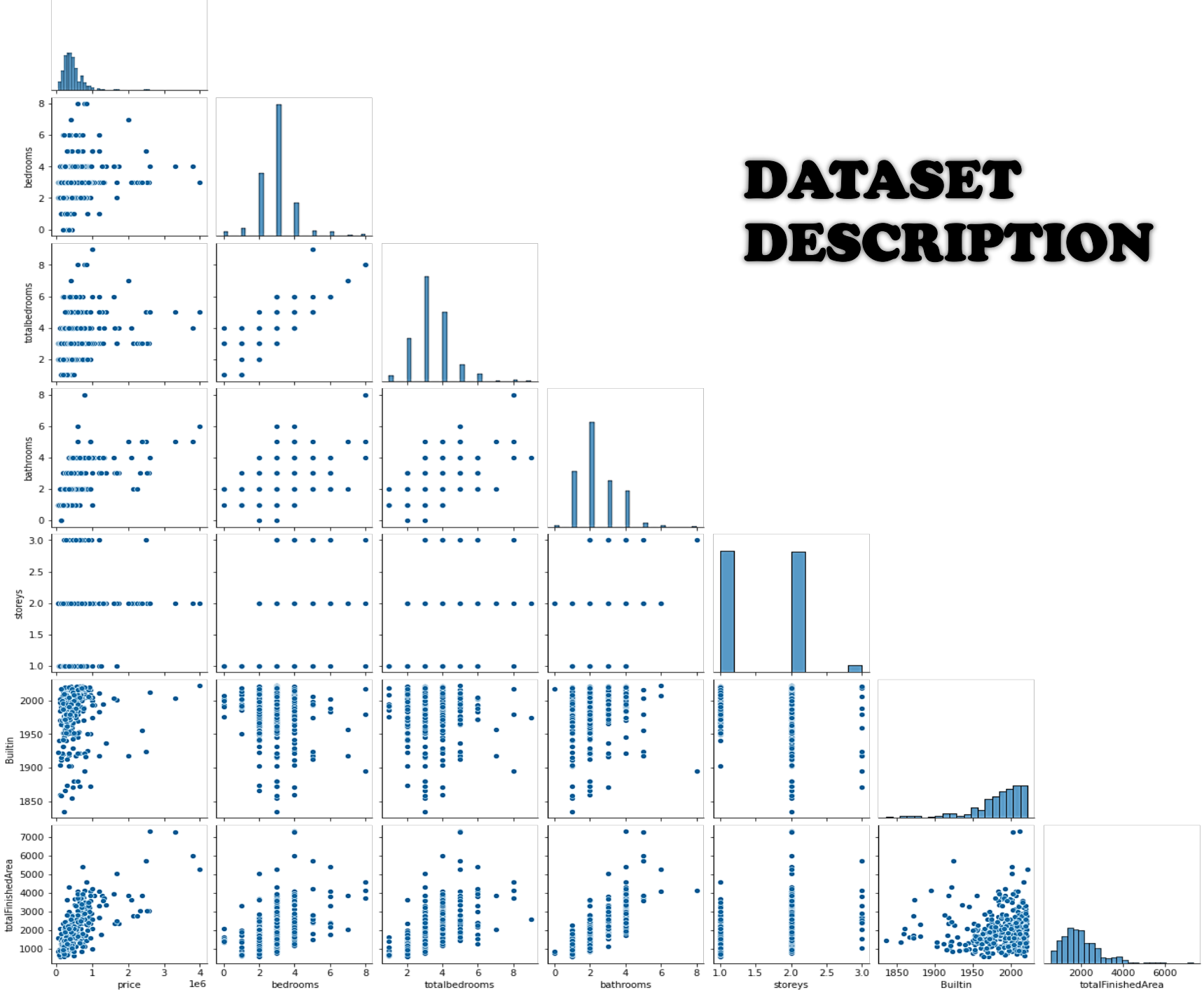
- Our Data Set is:
  - 510 instances
  - 20 Attributes
  - 7 numeric attributes
  - 13 nominal attributes



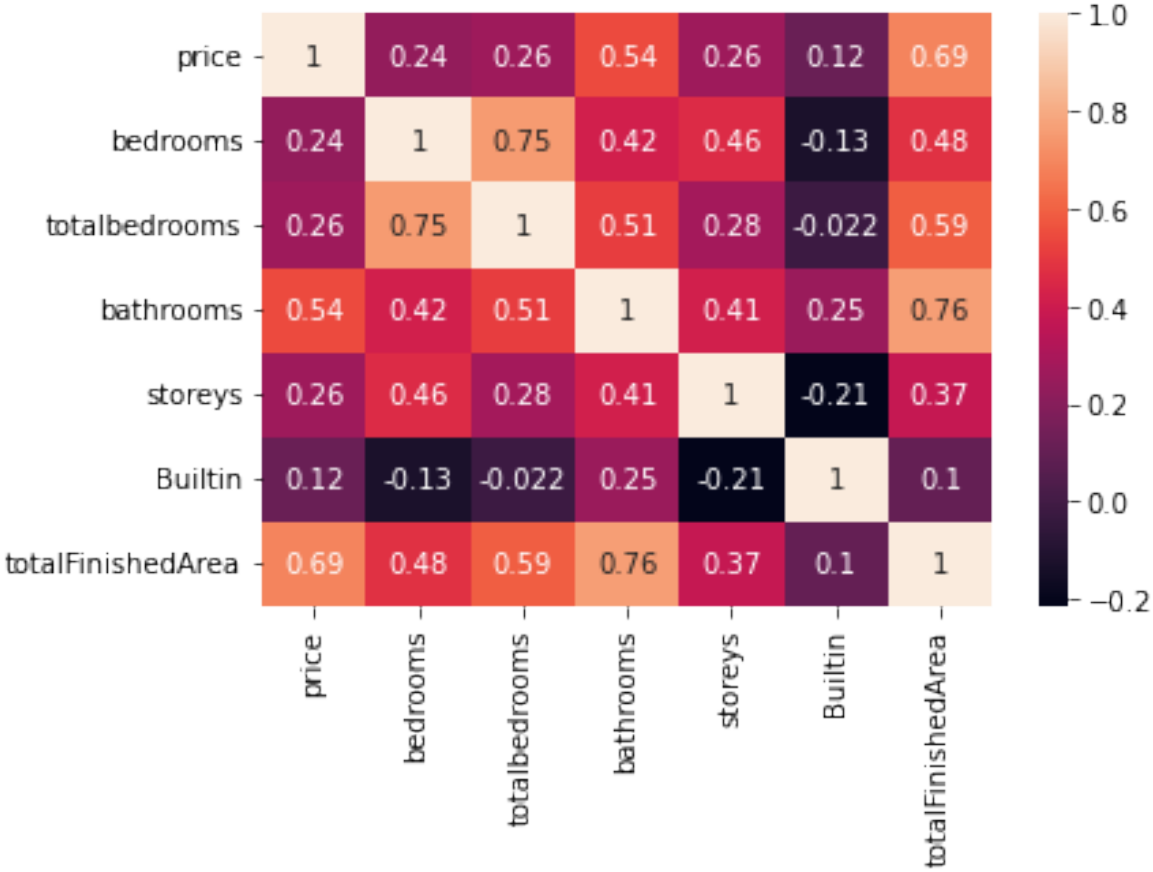
	propertyType	BuildingType	communityName	title	landSize	parkingType	appliancesIncluded	foundationType	style	architectureStyle	bas
0	Single Family	Row / Townhouse	Dartmouth	Freehold	under 1/2 acre	Garage	Stove, Dishwasher, Washer, Microwave Range Hoo...	Poured Concrete	NaN	NaN	
1	Single Family	Apartment	Dartmouth	Condominium/Strata	under 1/2 acre	Garage, Underground	Intercom	Poured Concrete	NaN	NaN	
2	Single Family	House	Kingston	Freehold	under 1/2 acre	NaN	Stove, Dryer, Washer, Microwave, Microwave Ran...	Poured Concrete	Semi-detached	3 Level	F
3	Single Family	House	Bedford	Freehold	under 1/2 acre	NaN	Stove, Dishwasher, Dryer, Washer, Refrigerator	Poured Concrete	Semi-detached	NaN	
4	Single Family	House	Robinsons Corner	Freehold	under 1/2 acre	Garage, Detached Garage	Cooktop - Electric, Oven, Dishwasher, Dryer, W...	Poured Concrete	Detached	NaN	

	price	bedrooms	totalbedrooms	bathrooms	storeys	Builtin	totalFinishedArea
0	399900	3	3	3	2	2010.0	1815.0
1	329900	2	2	1	1	2012.0	986.0
2	165000	3	4	2	2	1995.0	1400.0
3	339000	3	3	2	2	1980.0	1470.0
4	750000	3	4	3	2	2020.0	3120.0

DATASET DESCRIPTION

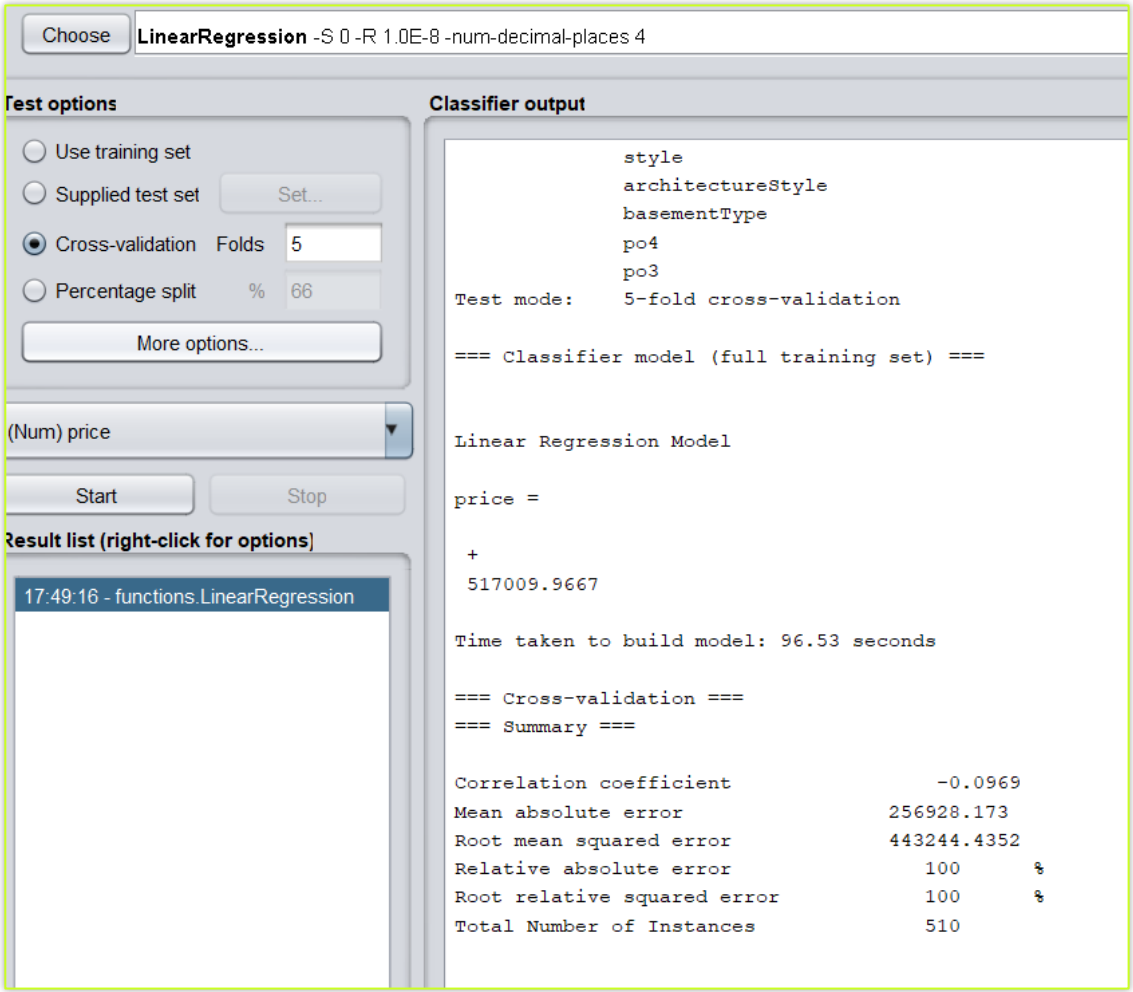


DATASET DESCRIPTION



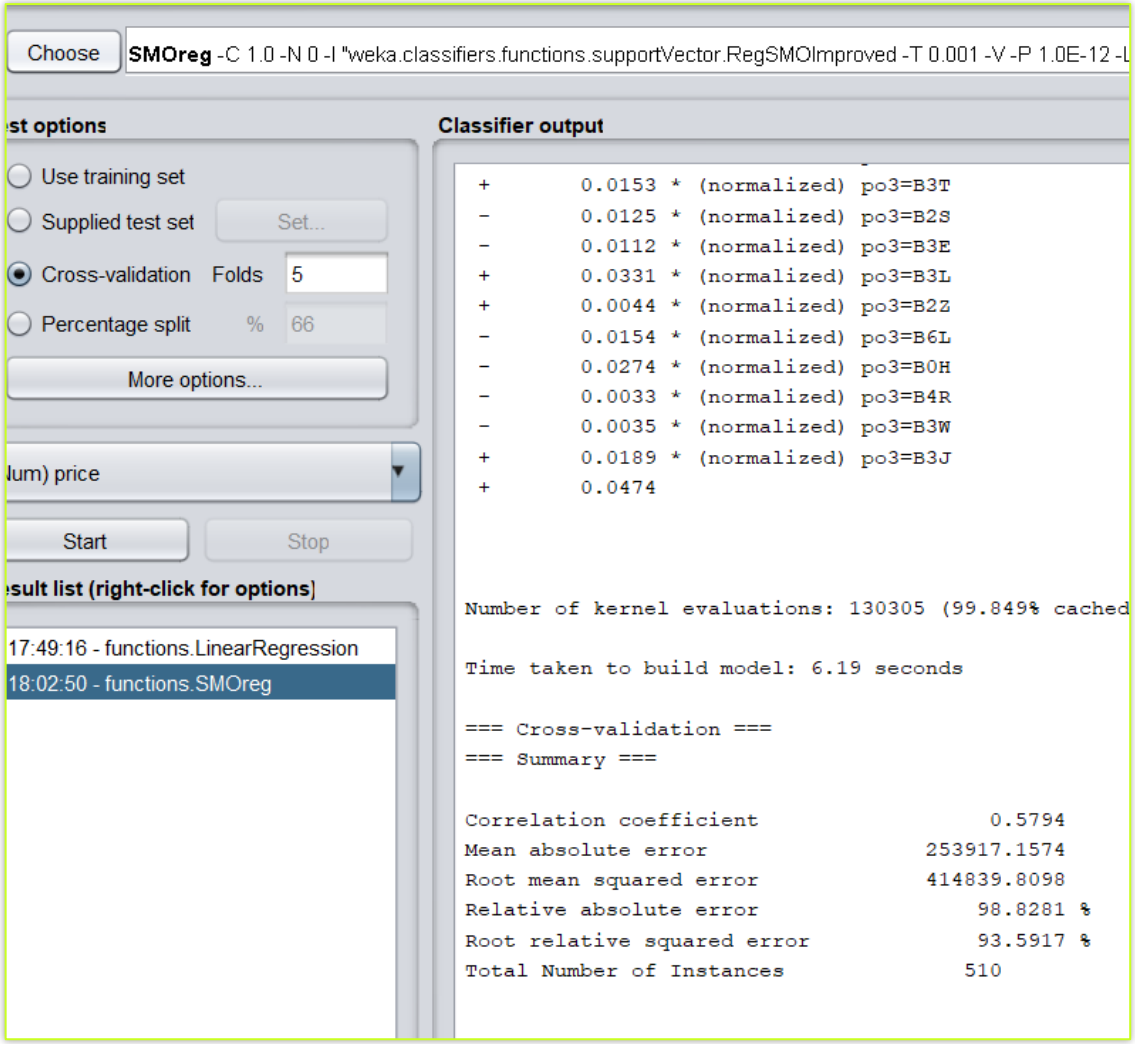
MODELS - 1

- Linear Regression
- Instances: 510
- Attributes: 20
- Test mode: 5-fold cross-validation
- Correlation coefficient -0.0969
- Mean absolute error 256928.173
- Root mean squared error 443244.4352
- Relative absolute error 100 %
- Root relative squared error 100 %



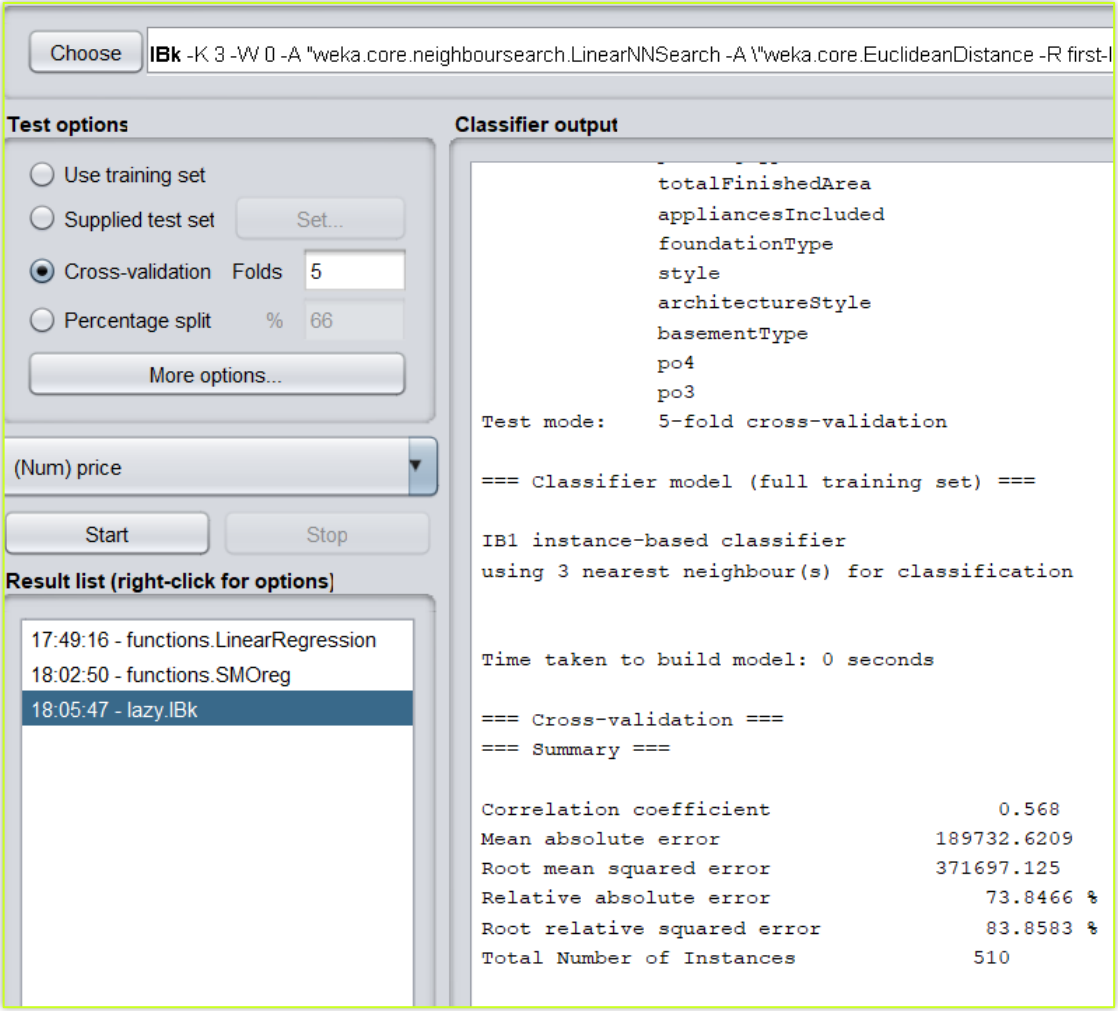
MODELS - 2

- SMO.Reg
- Instances: 510
- Attributes: 20
- Test mode: 5-fold cross-validation
- Correlation coefficient 0.5794
- Mean absolute error 253917.1574
- Root mean squared error 414839.8098
- Relative absolute error 98.8281 %
- Root relative squared error 93.5917 %



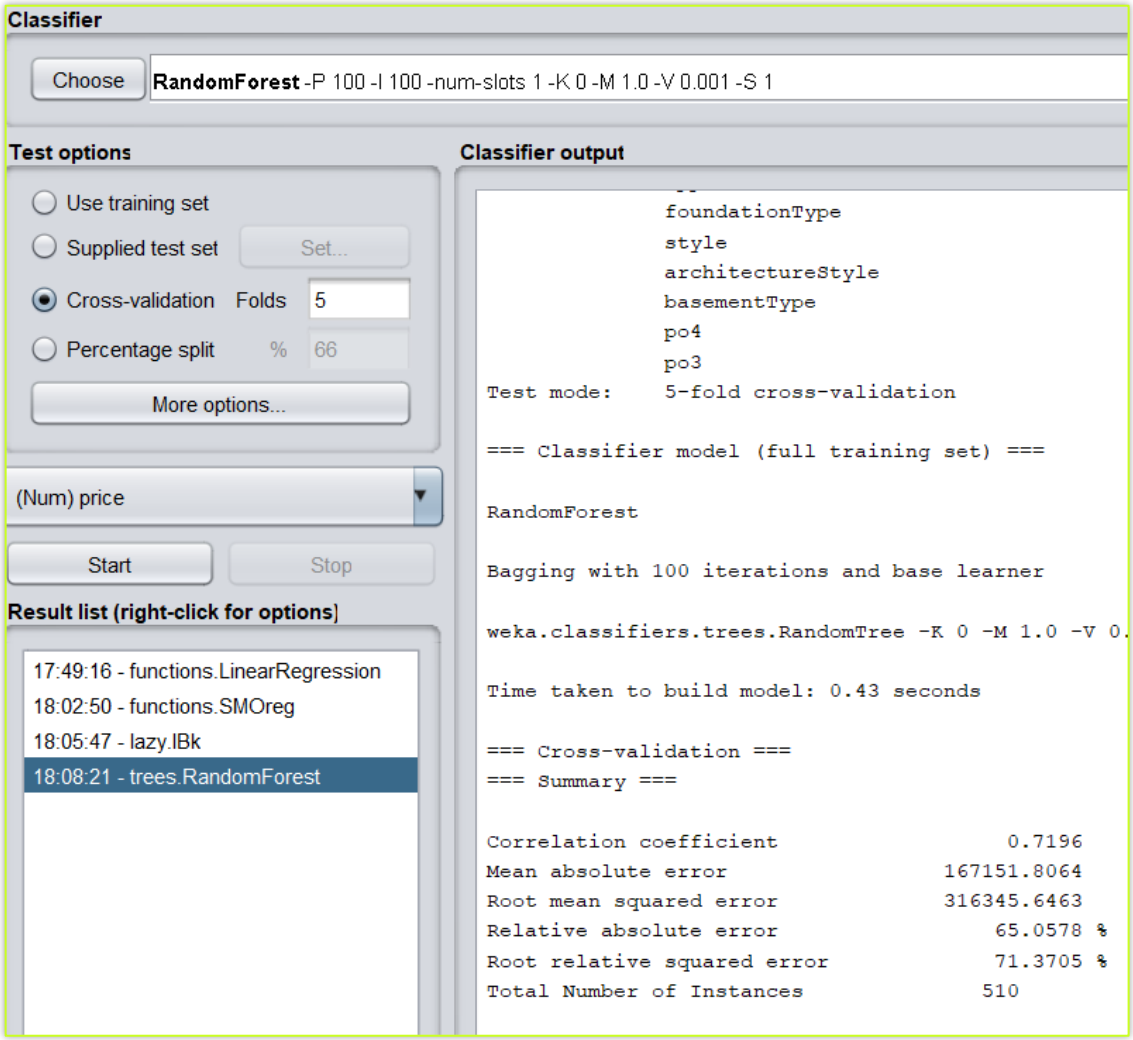
MODELS - 3

- K-nearest neighbors' classifier
- Instances: 510
- Attributes: 20
- Test mode: 5-fold cross-validation
- Correlation coefficient 0.568
- Mean absolute error 189732.6209
- Root mean squared error 371697.125
- Relative absolute error 73.8466 %
- Root relative squared error 83.8583 %



MODELS - 4

- Random Forest
- Instances: 510
- Attributes: 20
- Test mode: 5-fold cross-validation
- Correlation coefficient 0.7196
- Mean absolute error 167151.8064
- Root mean squared error 316345.6463
- Relative absolute error 65.0578 %
- Root relative squared error 71.3705 %



MODEL EVALUATION

Algorithm	Index	Base	After normalization	Removing imbalanced attributes 14 Attr.	Literature Review 7 Attr.
Random Forrest	Correlation cofiecient	0.72	0.72	0.72	0.72
	Root mean squared error	316345	0.08	0.08	0.08
	Relative squared error	71%	71%	72%	70%
Lazy.IBK	Correlation cofiecient	0.57	0.57	0.6	0.45
	Root mean squared error	371697	0.09	0.09	0.1
	Relative squared error	83%	82.40%	81%	90%
Linear Reg	Correlation cofiecient	-0.1	-0.1	0.72	0.74
	Root mean squared error	443244	0.11	0.08	0.08
	Relative squared error	100%	100%	73%	68%
SMO Reg	Correlation cofiecient	0.58	0.58	0.72	0.73
	Root mean squared error	414839	0.11	0.08	0.08
	Relative squared error	94%	94%	70%	71%

- Increasing the data
- Increasing in the geographical area
- Rental market
- Macro economic variables such as fluctuation
- Time serries
- Social media impact

FUTURE WORKS

Thank you for your patience

ANY QUESTIONS?