MIS772

Predictive Analytics

Workshop: Multiple Regression

Estimation, selecting attributes, coefficients, p-values, diagnostics and validation, and data pre-processing

Workshop Plan

Objectives:
Your task is to create a regression model to predict the sale price of houses in Aimes, USA. The aim is to improve the regression model performance by experimenting with pre-processing options.

Data Set:
Use files "AmesHousingPast.csv"

Acknowledgements:
Dean De Cock, Truman State University, 2011.

Original Data from Kaggle:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Method:
Attend the seminar, follow the tutor's demo and instructions, take notes. Note that the class and online seminar will be recorded and their videos linked to the CloudDeakin topic for later access and study.

DEAKIN BUSINESS SCHOOL

Acquire data for estimation

- Load Ames housing data and unzip
- Chart and explore past data

Investigate correlation of attributes

- Select numeric attributes only, set SalePrice as a label
- Explore correlation tables, take notes
- Explore "Weight by Correlation" and its results

Create a simple regression model

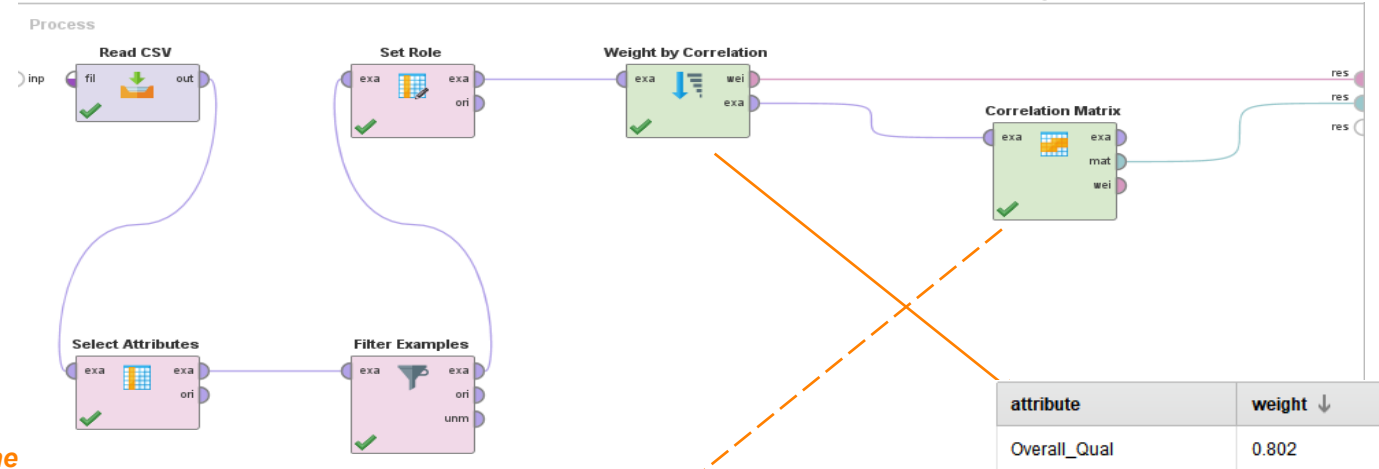
- Explore "Lot_Frontage" vs "SalePrice"
- Create a simple cross-validated regression model
- Run and take notes of performance!

Create a multiple regression model

- In steps: change, run, check performance against notes
- Modify previous process to include all numeric attrs
- Experiment with attribute selection
- Experiment with regression collinearity option
- Calculate residuals
- Reflect on coefficients, p-values and tolerance
- *Optional: Dummy encode nominal attributes*

Results: Correlation

We will start by loading AmesHousing (past) data, taking its sample if necessary (50%), selecting numeric attributes only, selecting SalePrice as a label, weighing attributes by correlation and creating a correlation matrix. Then investigate the results.



Also check the correlation matrix visualisation (in colour)

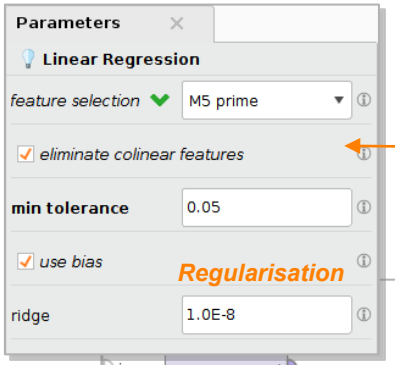
What does it mean?

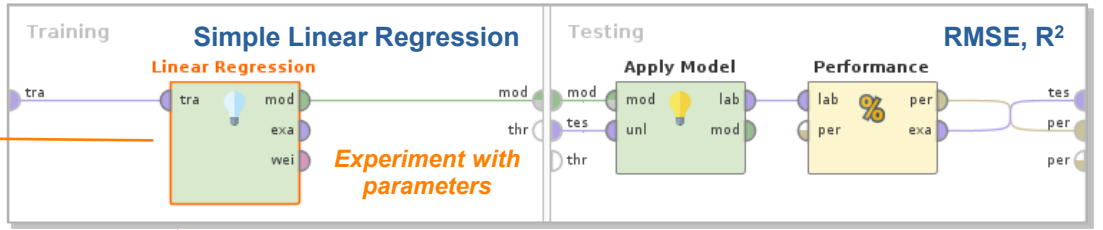
Attributes	PID	MS_S...	Lot_F...	Lot_A...	Over...	Over...	Mas_...	Bsmt...
PID	1	-0.000	-0.096	0.039	-0.259	0.108	-0.228	-0.089
MS_SubClass	-0.000	1	-0.415	-0.194	0.040	-0.056	0.009	-0.052
Lot_Frontage	-0.096	-0.415	1	0.489	0.225	-0.079	0.233	0.224
Lot_Area	0.039	-0.194	0.489	1	0.095	-0.038	0.127	0.194
Overall_Qual	-0.259	0.040	0.225	0.095	1	-0.101	0.442	0.290
Overall_Cond	0.108	-0.056	-0.079	-0.038	-0.101	1	-0.140	-0.059
Mas_Vnr_Area	-0.228	0.009	0.233	0.127	0.442	-0.140	1	0.309
BsmtFin_SF_1	-0.089	-0.052	0.224	0.194	0.290	-0.059	0.309	1
BsmtFin_SF_2	-0.003	-0.065	0.046	0.087	-0.037	0.028	-0.013	-0.047
Bsmt_Unf_SF	-0.094	-0.130	0.115	0.021	0.272	-0.127	0.087	-0.478
Total_Bsmt_SF	-0.186	-0.207	0.360	0.254	0.555	-0.177	0.401	0.539
1st_Flr_SF	-0.134	-0.235	0.464	0.329	0.482	-0.165	0.400	0.461

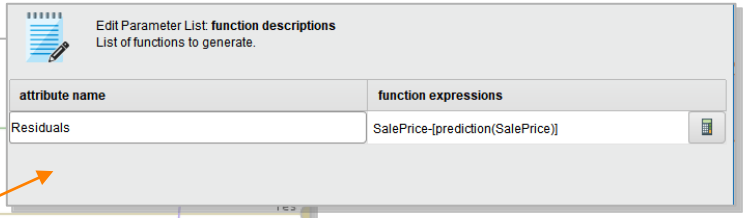
attribute	weight ↓
Overall_Qual	0.802
Gr_Liv_Area	0.709
Garage_Cars	0.648
Garage_Area	0.643
Total_Bsmt_SF	0.629
1st_Flr_SF	0.619
Year_Built	0.555
Full_Bath	0.553
Year_Remod/Add	0.535
Garage_Yr_Blt	0.524

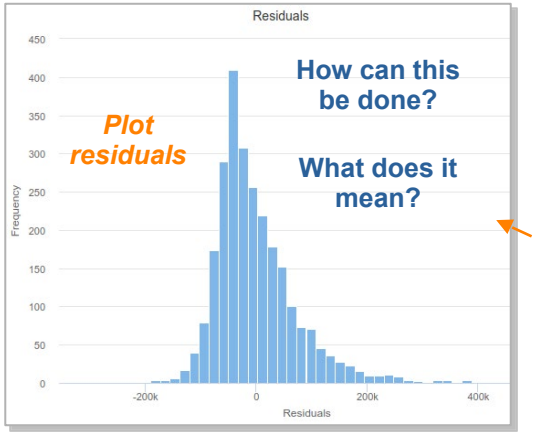
DEAKIN BUSINESS SCHOOL

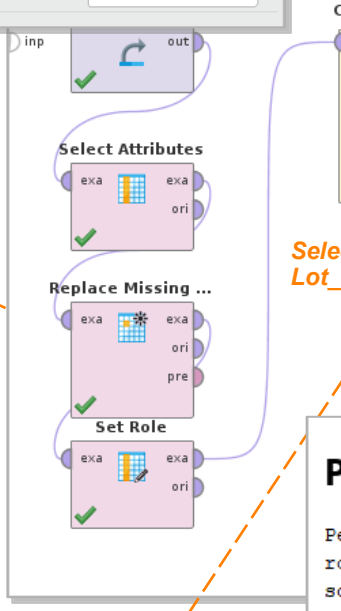
Simple Regression

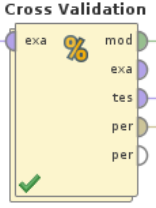


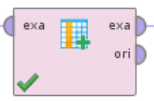












Select attributes: Lot_Frontage + SalePrice

Calculate residuals

Run

What does it mean?

Observe changing performance as you experiment with regression parameters

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 78332.206 +/- 1406.054 (micro average: 78340.155 +/- 0.000)
squared_correlation: 0.134 +/- 0.025 (micro average: 0.131)
```

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
Lot_Frontage	1307.497	69.236	0.348	1	18.885	0	****
(Intercept)	91350.549	5017.878	?	?	18.205	0	****

What does it mean?

DEAKIN BUSINESS SCHOOL

Multiple Linear Regression

- How would you change the previous process to include multiple predictors?
- Compare the performance of the simple vs the multiple regression model.
- Check that the multiple regression model meets the assumptions of linear regression:
  - Multicollinearity (which attributes are best to exclude?)
  - Residuals
- Why is “throwing all input attributes” into a linear predictive model bad modelling practice?