MIS772
Predictive Analytics

Linear Regression Models as estimation method

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

Multiple regression

- **Understanding a linear model**
  coefficients, p-values, R^2
- **The fundamental assumptions of regression modelling**
- **Model diagnostics**
- **Attribute selection**

Case Study: Ames Real Estate

Ames Real Estate Data Set (source: kaggle.com):

79 regular attributes describing (almost) every aspect of residential homes in Ames, Iowa, US

This competition challenges you to predict the label attribute, i.e. SALE PRICE of each home.

ExampleSet (2930 examples, 2 special attributes, 79 regular attributes)

Row No.	PID	SalePrice	MS_SubClass	MS_Zoning	Lot_Frontage	Lot_Area	Street	Alley	Lot_Shape	Land_Contour	Utilities	Lot_Config
1	526301100	215000	20	RL	141	31770	Pave	NA	IR1	Lvl	AllPub	Corner
2	526350040	105000	20	RH	80	11622	Pave	NA	Reg	Lvl	AllPub	Inside
3	526351010	172000	20	RL	81	14267	Pave	NA	IR1	Lvl	AllPub	Corner
4	526353030	244000	20	RL	93	11160	Pave	NA	Reg	Lvl	AllPub	Corner
5	527105010	189900	60	RL	74	13830	Pave	NA	IR1	Lvl	AllPub	Inside
6	527105030	195500	60	RL	78	9978	Pave	NA	IR1	Lvl	AllPub	Inside
7	527127150	213500	120	RL	41	4920	Pave	NA	Reg	Lvl	AllPub	Inside
8	527145080	191500	120	RL	43	5005	Pave	NA	IR1	HLS	AllPub	Inside
9	527146030	236500	120	RL	39	5389	Pave	NA	IR1	Lvl	AllPub	Inside
10	527162130	189000	60	RL	60	7500	Pave	NA	Reg	Lvl	AllPub	Inside
11	527163010	175900	60	RL	75	10000	Pave	NA	IR1	Lvl	AllPub	Corner
12	527165230	185000	20	RL	?	7980	Pave	NA	IR1	Lvl	AllPub	Inside

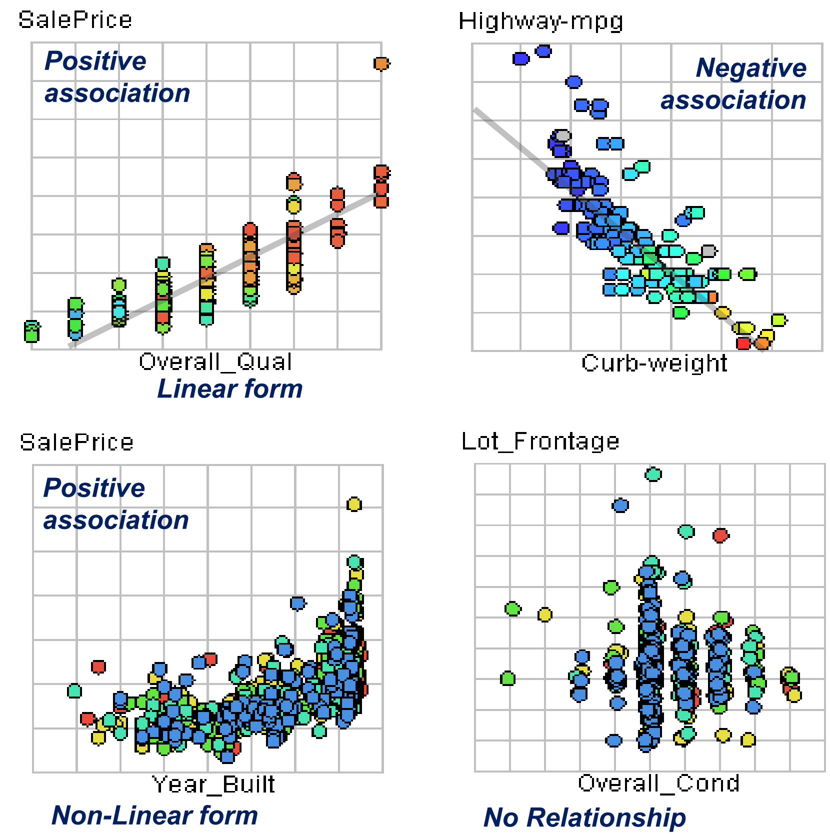
We can build a predictive model for the **Sale Price** based on other attributes (predictors) available in the data set, thanks to the **relationships** existing between them.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

DEAKIN BUSINESS SCHOOL

Relationships between numeric attributes

- There exist many kinds of ***relationships*** between attributes.
- One such relationship is ***correlation***.
- Attributes are correlated when an increase of value in one attribute is accompanied by the simultaneous increase (top left) or decrease (top right) in the value of another.
- The rate of such increase could indicate linear (top row) or non-linear dependency.
- Correlation does not imply ***causation***, which indicates that changes to values of one attribute are in direct consequence of changes in another.
- Correlation can be described in terms of ***association***, ***form*** and ***strength***.
- Scatter plots are useful in visual identification of correlation in small data sets.
- In large data sets, scatter plots are very confusing and have little value.



Selected charts from Houses and Cars data sets

Also see KD 3.3 on Correlation

DEAKIN BUSINESS SCHOOL

Correlation

- The most common measure of correlation is ***Pearson's correlation***, which measures linear dependency between numerical variables.
- Pearson's correlation coefficient for two ***normally distributed variables*** indicates that if one variable's value increases the other also changes values ***consistently***.
- Correlation is positive when values of two variables fluctuate together and grow in the same direction.
- Correlation of +1 indicates identity (x=x).
- Correlation is negative when two variables fluctuate together, but values of one variable grows while the other decreases.
- Correlation of -1 indicates -identity (x=-x).
- Correlation is close to zero when there is little relationship between two variables.
- Correlation of 0 indicates completely random pairing of variables' values.

The following example shows a correlation table of house attributes

Attributes	PID	MS_Su...	Lot_Fro...	Lot_Area	Overall...	Overall...	Year_B...	Year_R...	Mas_Vn...	BsmtFin...	BsmtFin...	Bsmt_U...	Total_B...	1st_Flr_...	2nd_Flr...
					-0.263	0.104	-0.343	-0.157	-0.229	-0.098	-0.001	-0.088	-0.190	-0.142	-0.003
					0.039	-0.067	0.037	0.043	0.003	-0.060	-0.071	-0.130	-0.219	-0.248	0.304
Lot_Frontage	-0.097	-0.420	1	0.491	0.212	-0.074	0.122	0.092	0.222	0.216	0.046	0.117	0.354	0.457	0.029
Lot_Area	0.035	-0.205	0.491	1	0.097	-0.035	0.023	0.022	0.127	0.192	0.083	0.024	0.254	0.332	0.033
Overall_Qual	-0.263	0.039	0.212	0.097	1	-0.095	0.597	0.570	0.429			0.270	0.547	0.478	0.241
Overall_Cond	0.104	-0.067	-0.074	-0.035	-0.095	1	-0.369	0.048	-0.135			-0.137	-0.173	-0.157	0.006
Year_Built	-0.343	0.037	0.122	0.023	0.597	-0.369	1	0.612	0.313	0.280	-0.027	0.129	0.408	0.310	0.017
Year_Remod/Add	-0.157	0.043	0.092	0.022	0.570	0.048	0.612	1	0.197	0.152	-0.062	0.165	0.297	0.242	0.159
Mas_Vnr_Area	-0.229	0.003	0.222	0.127	0.429	-0.135	0.313	0.197	1	0.302	-0.016	0.092	0.397		
BsmtFin_SF_1	-0.098	-0.060			0.284	-0.051	0.280	0.152	0.302	1	-0.054	-0.478	0.537		
BsmtFin_SF_2	-0.001	-0.071			-0.041	0.041	-0.027	-0.062	-0.016	-0.054	1	-0.239	0.090	0.085	-0.098
Bsmt_Unf_SF	-0.088	-0.130			0.270	-0.137			0.092			1	0.412	0.296	0.002
Total_Bsmt_SF	-0.190	-0.219	0.354	0.254	0.547	-0.173	0.408	0.297	0.397			0.412	1	0.801	-0.205
					0.478	-0.157	0.310	0.242	0.396	0.457	0.085	0.296	0.801	1	-0.250
					0.241	0.006	0.017	0.159	0.122	-0.164	-0.098	0.002	-0.205	-0.250	1

Little or no correlation

Moderately high correlation

Identity

Very high correlation

Negative correlation

What correlation is high or low depends on application and its data

Some commonly used types of correlation coefficients:
- *Pearson r* (linear relationship, assumes normal distribution, sensitive to outliers)
- *Spearman rho* (monotonic relationship, non-parametric, based on deviations, not linear, no normality)
- *Kendall tau* (between any ordered vars, non-parametric, based on concordance - same order, good for small samples)

DEAKIN BUSINESS SCHOOL

Linear Models

- Relationships between two variables can often be approximated by an equation describing their linear combination, which defines a *linear model*.

- In case of two variables, the equation describes a line, which is referred to as a *regression line or simple regression.*

$$y = b_0 + b_1 \times x + e$$

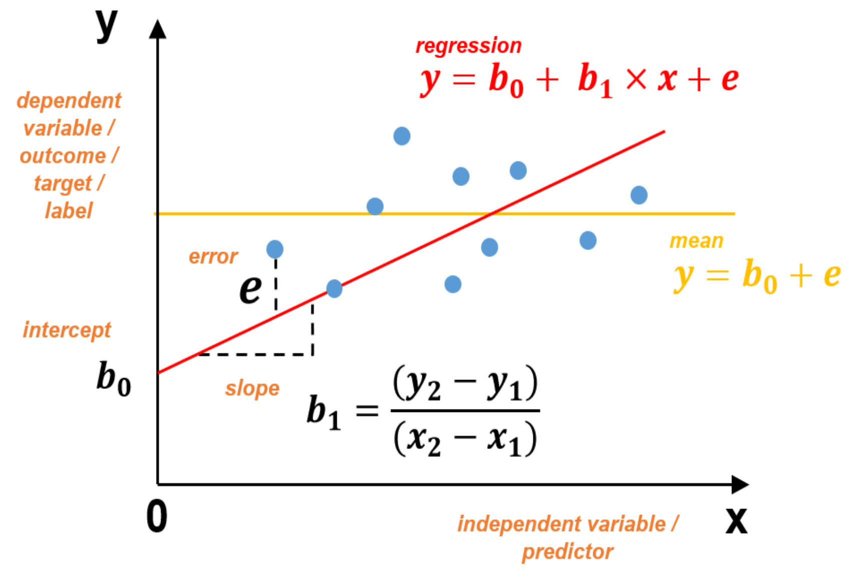
- The regression line can be defined by a mathematical formula for a line, defined by its *intercept* with the axis of the outcome variable (where x=0), its *slope* (proportion between x and y) and *error term*.

- When we have more variables we describe a *multiple regression*.

$$y = b_0 + b_1 \times x_1 + b_2 \times x_2 + \cdots b_n \times x_n + e$$

- Regression analysis is the most commonly used predictive model.

- When predicted values are calculated using the regression formula, the total error, or differences between the expected values and the actual values gives an idea as to the model quality.



- The following measures are commonly used:
  - *MAE* (mean absolute error),
  - *RMSE* (root mean square error)

- Regression also assesses the quality of its predictions. It calculates a metric which indicates how much variance in the label attribute the model can explain from the input predictors, this is called the *coefficient of determination*:
  - R^2 (coefficient of determination).

Also see KD 5.1 on Intro to regression

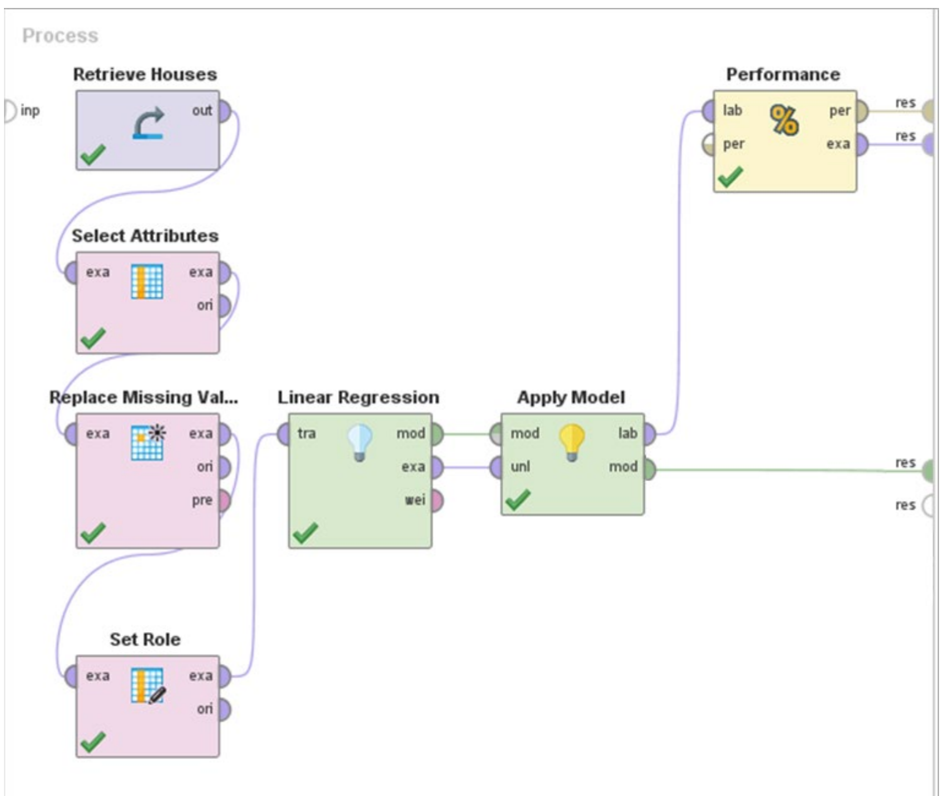
DEAKIN BUSINESS SCHOOL

Linear Regression Modelling

Your task is: *Create an **estimator** (estimation model) to predict the house price in Ames*

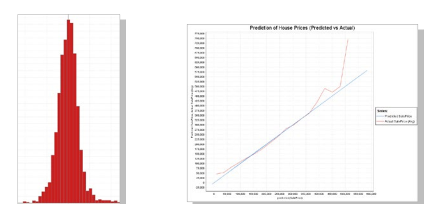
Plan of action:

- Acquire data
- Study variables involved
- Create a simple regression model
- Create a multiple regression model
- Create and interpret various regression diagnostic charts



Regression model's assumptions:

- All variables are numeric
- No missing/bad values
- No extreme cases
- All predictors are independent (no multi-collinearities)
- Prediction errors (residuals) are normally distributed



There are several diagnostic plots of ensuring regression model quality (RapidMiner)

DEAKIN BUSINESS SCHOOL

Pros and Cons of Regression Modelling

Pros

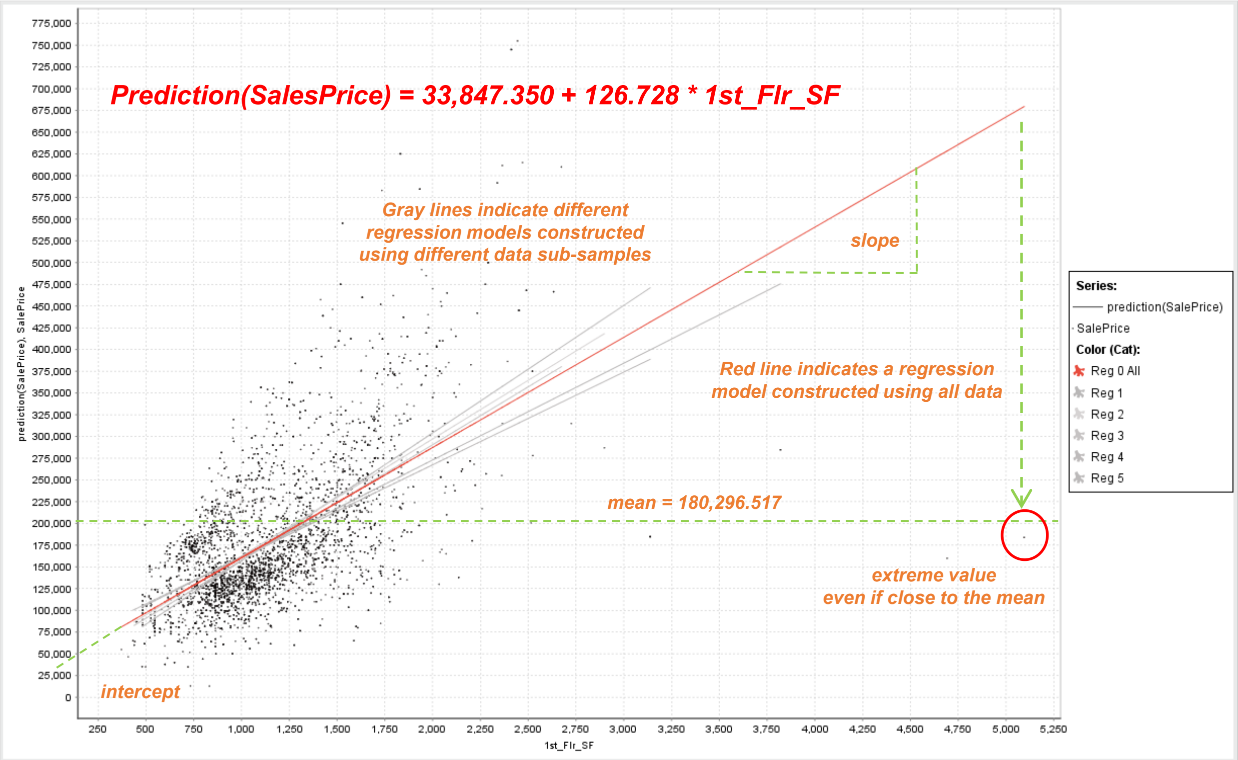
- Most common approach for modelling numeric data
- Can be (and is being) adapted to model almost any data
- Provides estimates of the strength and size of the relationships among independent variables and a dependent variable
- Can be visualised (not easy in case of multiple regression)

Cons

- Makes strong assumptions about data
- The model's form must be specified by the user in advance
- Cannot handle missing data
- Only works with numeric features, so categorical attributes may need dummy encoding

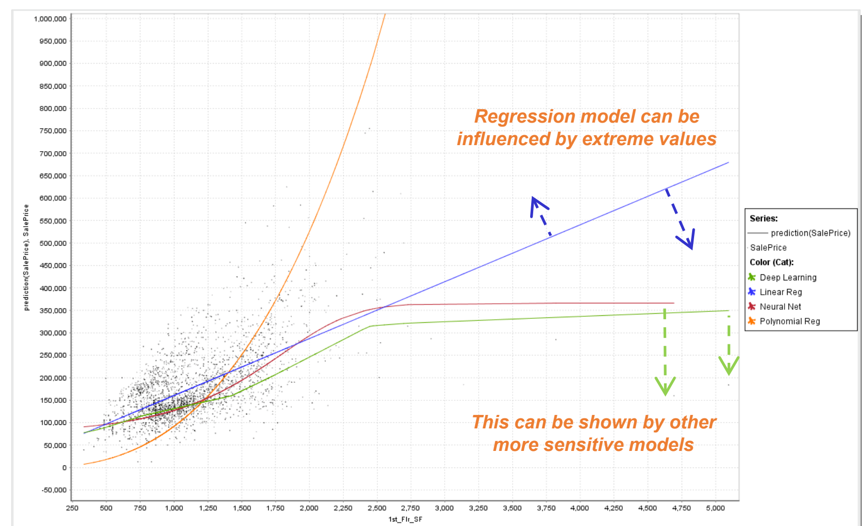
DEAKIN BUSINESS SCHOOL

Linear Regression



Example:
A scatter plot showing the relationship btw. the floor area vs the house price
A training sample defines a regression line, which could significantly vary! (gray lines - left)
Cross-validation (or bootstrap validation) is thus very important!

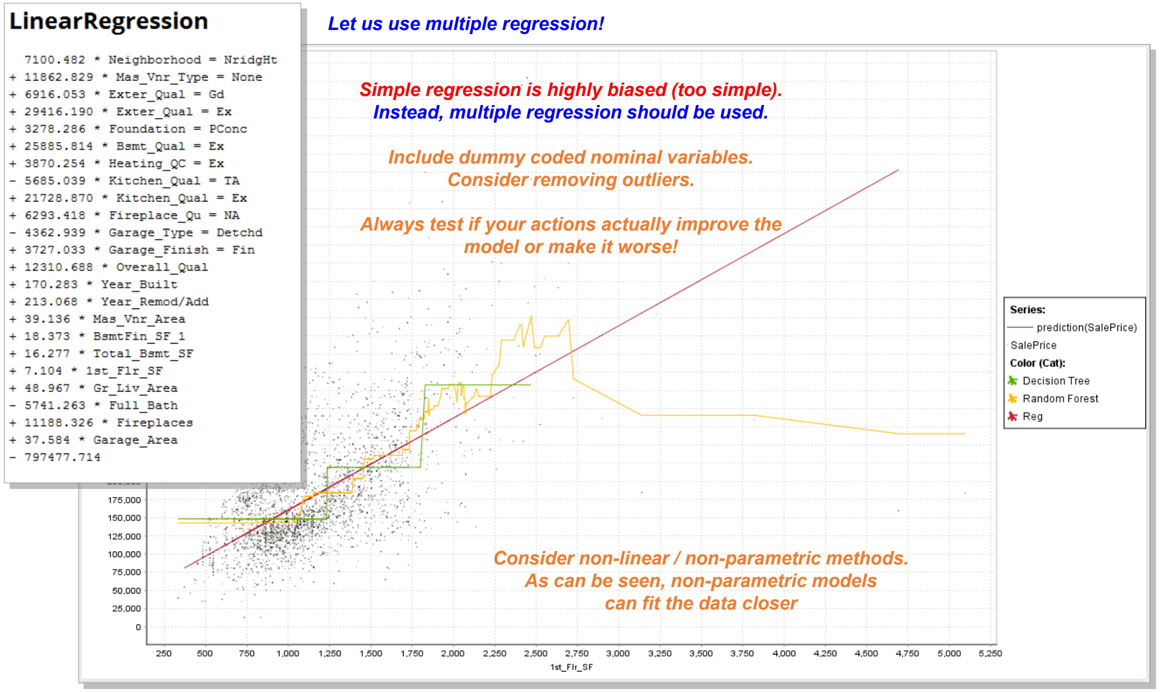
- **Large houses cost more: given the floor surface area we can estimate the price**
- **Visually, the relationship appears to have a linear trend, which may suggest using a linear model to estimate sales price as a function of floor surface area**
- **And yet, on closer look – by applying more sensitive models, we can find that extreme values (i.e. outliers) in data could distort this view significantly**



We need to eliminate outliers

DEAKIN BUSINESS SCHOOL

Linear Regression



- When we deal with many predictors, we need to create a ***multiple regression model***
- It is important that all predictors are independent
- ***Multi-collinearity of attributes*** implies that predictors are actually closely related (i.e. one predictor can be estimated from the other predictors). This is problematic predictive modelling.
- The seminars this week will illustrate how multi-collinearity can be identified in RapidMiner.
- RapidMiner can eliminate attribute multi-collinearities from multiple regression
- There are also many methods of ***attribute selection*** for regression, e.g. greedy or M5 Prime

Improving Performance of linear models:

Consider:

- metrics (R^{2}, RMSE, etc.)
- The business context
- Logic of why a predictor(s) would influence the predicted attribute
- Parsimonious (less is more!) models are preferred in many business contexts

DEAKIN BUSINESS SCHOOL

Multiple Regression: Coefficients Analysis

Analysis of Regression Coefficients

- **The coefficients represent the amount of change to expect in the label if there was a one-unit change in the predictor attribute.**
- **The largest positive coefficients are "Overall_Qual" and "Overall_Cond" → better quality / condition, higher the price**
- ***Only a few attributes influence the price***

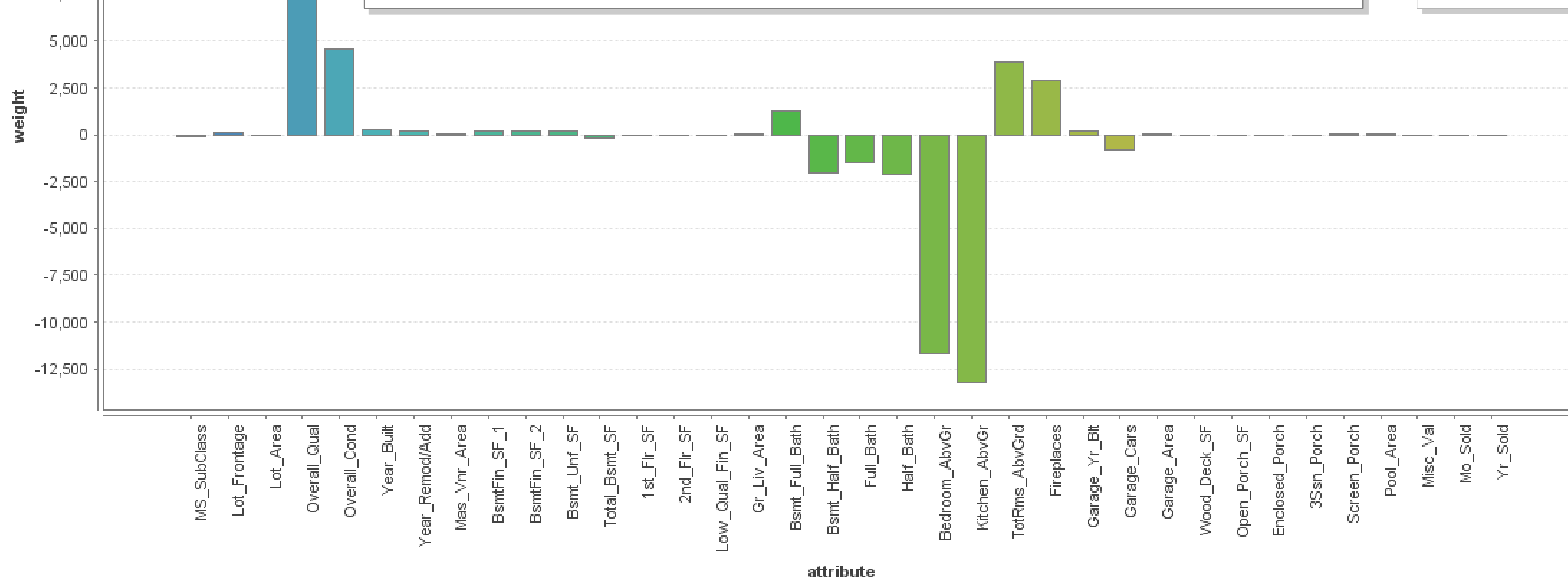
LinearRegression

```
- 101.898 * MS_SubClass
+ 147.582 * Lot_Frontage
+ 0.576 * Lot_Area
+ 15301.046 * Overall_Qual
+ 4606.237 * Overall_Cond
+ 273.158 * Year_Built
+ 199.149 * Year_Remod/Add
+ 27.907 * Mas_Vnr_Area
+ 223.492 * BsmtFin_SF_1
+ 203.531 * BsmtFin_SF_2
+ 202.005 * Bsmt_Unf_SF
- 177.893 * Total_Bsmt_SF
- 26.764 * 1st_Flr_SF
- 23.947 * 2nd_Flr_SF
- 47.517 * Low_Qual_Fin_SF
+ 85.737 * Gr_Liv_Area
+ 1297.416 * Bsmt_Full_Bath
- 1997.572 * Bsmt_Half_Bath
- 1513.752 * Full_Bath
- 2092.151 * Half_Bath
- 11686.019 * Bedroom_AbvGr
- 13209.935 * Kitchen_AbvGr
+ 3844.733 * TotRms_AbvGrd
+ 2922.100 * Fireplaces
+ 164.406 * Garage_Yr_Blt
- 788.296 * Garage_Cars
+ 25.584 * Garage_Area
+ 34.525 * Screen_Porch
+ 23.592 * Pool_Area
- 1321556.933
```

It should be noted that the size of the coefficients depends on the units of attributes passed into the regression - their size will be adjusted to compensate for the range of the variable values and the relative difference in variable units

Strange positive / negative impact of attributes on the label may hint on unusual coding or interpretation of data.

It could also mean that for the purpose of coefficient analysis attributes should be normalised!



DEAKIN BUSINESS SCHOOL

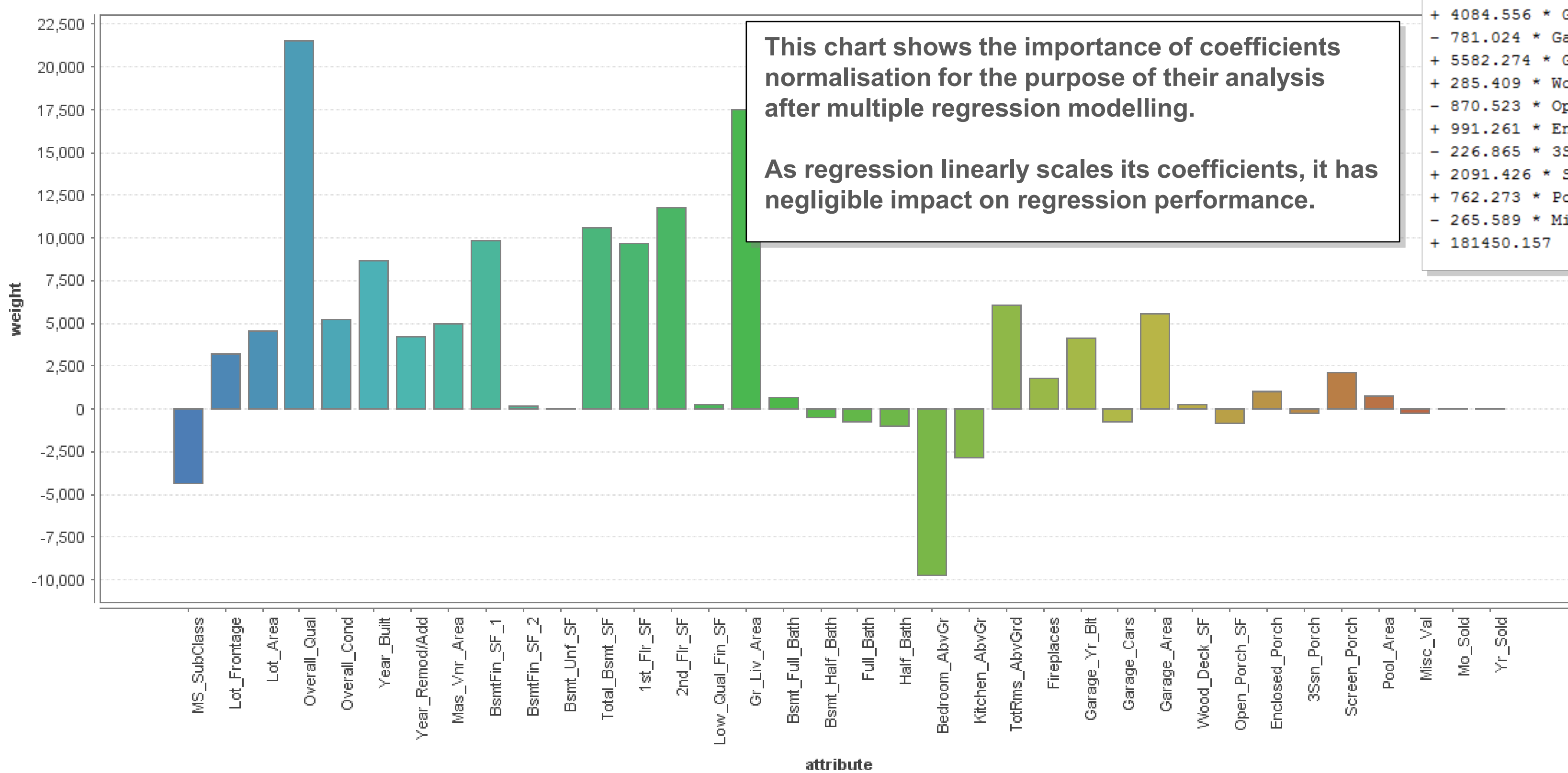
Multiple Regression: Coefficients Analysis

Analysis of Regression Coefficients

- **To analyse the impact of regression coefficients, attributes should be initially assumed to be of equal importance**
- **This can be done by normalising their values (e.g. Z-transform)**
- **This has negligible impact on the model performance!**
- **The largest positive coefficients "Overall_Qual" and "Gr_Liv_Area" → better overall quality and larger living area, higher the price**
- **The figure shows that the <u>price is influenced by many factors</u>**

LinearRegression

```
- 4375.632 * MS_SubClass
+ 3174.601 * Lot_Frontage
+ 4525.996 * Lot_Area
+ 21517.558 * Overall_Qual
+ 5193.555 * Overall_Cond
+ 8689.946 * Year_Built
+ 4230.949 * Year_Remod/Add
+ 4997.577 * Mas_Vnr_Area
+ 9818.219 * BsmtFin_SF_1
+ 184.897 * BsmtFin_SF_2
+ 10619.638 * Total_Bsmt_SF
+ 9692.426 * 1st_Flr_SF
+ 11769.446 * 2nd_Flr_SF
+ 218.693 * Low_Qual_Fin_SF
+ 17521.622 * Gr_Liv_Area
+ 676.707 * Bsmt_Full_Bath
- 463.973 * Bsmt_Half_Bath
- 740.567 * Full_Bath
- 990.634 * Half_Bath
- 9736.961 * Bedroom_AbvGr
- 2818.532 * Kitchen_AbvGr
+ 6036.201 * TotRms_AbvGrd
+ 1790.425 * Fireplaces
+ 4084.556 * Garage_Yr_Blt
- 781.024 * Garage_Cars
+ 5582.274 * Garage_Area
+ 285.409 * Wood_Deck_SF
- 870.523 * Open_Porch_SF
+ 991.261 * Enclosed_Porch
- 226.865 * 3Ssn_Porch
+ 2091.426 * Screen_Porch
+ 762.273 * Pool_Area
- 265.589 * Misc_Val
+ 181450.157
```



This chart shows the importance of coefficients normalisation for the purpose of their analysis after multiple regression modelling.

As regression linearly scales its coefficients, it has negligible impact on regression performance.

DEAKIN BUSINESS SCHOOL

Diagnostic Charts vs Goodness of Fit

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Neighborhood = NridgHt	5030.569	2705.756	0.015	0.790	1.859	0.063	*
Mas_Vnr_Type = None	10949.336	1562.229	0.067	0.793	7.009	0.000	****
Exter_Qual = Gd	7920.205	5417.798	0.047	0.777	1.462	0.144	
Exter_Qual = TA	3397.221	5079.672	0.021	0.601	0.669	0.504	
Exter_Qual = Ex	34154.604	6636.864	0.080	0.752	5.146	0.000	****
Foundation = PConc	2205.366	1663.621	0.014	0.701	1.326	0.185	
Bsmt_Qual = Ex	26953.961	2507.857	0.097	0.651	10.748	0	****
Bsmt_Qual = TA	-867.878	1539.929	-0.005	0.776	-0.564	0.573	
Heating_QC = Ex	4705.045	1327.494	0.029	0.779	3.544	0.000	****
Kitchen_Qual = Ex	24560.037	2772.996	0.080	0.710	8.857	0	****
Kitchen_Qual = TA	-5570.479	1524.418	-0.035	0.700	-3.654	0.000	****
Fireplace_Qu = Gd	3896.583	1545.896	0.021	0.845	2.521	0.012	**
Fireplace_Qu = NA	5644.848	2558.791	0.035	0.732	2.206	0.027	**
Garage_Finish = Fin	3528.981	1471.162	0.019	0.808	2.399	0.017	**
Garage_Finish = Unf	-1189.953	1431.747	-0.007	0.800	-0.831	0.406	
Overall_Qual	10564.290	688.659	0.187	0.351	15.340	0	****
Mas_Vnr_Area	38.437	4.410	0.085	0.740	8.716	0	****
BsmtFin_SF_1	21.585	1.459	0.119	0.846	14.793	0	****
Total_Bsmt_SF	15.672	2.346	0.083	0.547	6.681	0.000	****
1st_Flr_SF	7.683	2.585	0.036	0.548	2.972	0.003	***
Gr_Liv_Area	59.985	2.430	0.370	0.488	24.681	0	****
Full_Bath	-4120.002	1432.609	-0.028	0.634	-2.876	0.004	***
TotRms_AbvGrd	-2177.602	596.085	-0.042	0.696	-3.653	0.000	****
Fireplaces	8960.712	1944.569	0.072	0.761	4.608	0.000	****
Garage_Cars	906.981	1673.757	0.009	0.529	0.542	0.588	
Garage_Area	27.067	5.670	0.073	0.550	4.774	0.000	****
Year_Built	225.608	34.598	0.086	0.673	6.521	0.000	****
Year_Remod/Add	209.689	37.723	0.055	0.703	5.559	0.000	****
Garage_Yr_Blt	-24.679	38.042	-0.008	0.700	-0.649	0.517	
(Intercept)	-848706.078	92466.655	?	?	-9.179	0	****

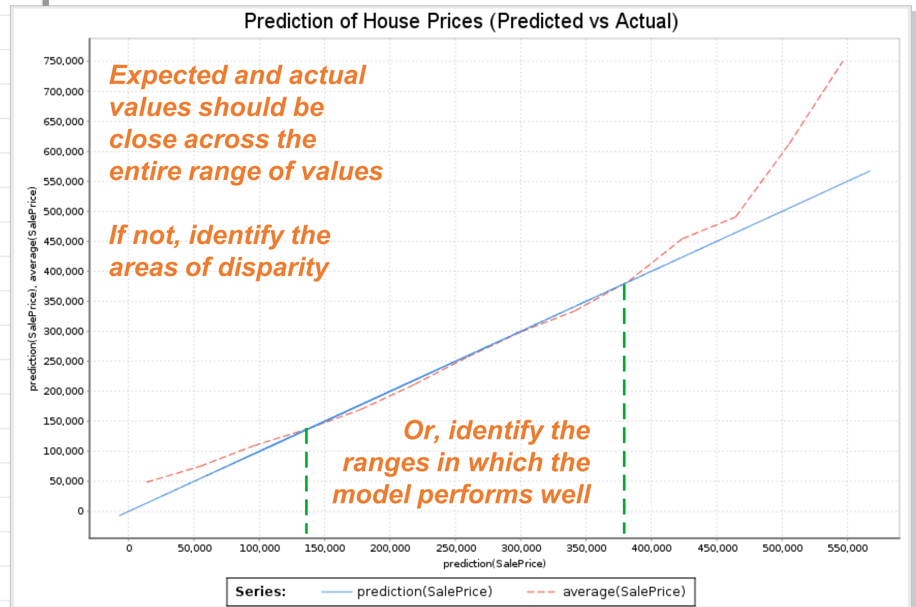
Coefficients should have small p-values < 0.05

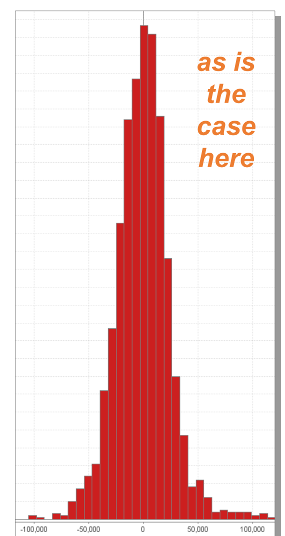
Not always – it depends on var selection methods!

PerformanceVector

Low RMSE and High R^2 (from Cross-Validation)

```
PerformanceVector:
root_mean_squared_error: 26850.937 +/- 7817.039 (mikro: 27960.905 +/- 0.000)
correlation: 0.939 +/- 0.028 (mikro: 0.935)
squared_correlation: 0.883 +/- 0.051 (mikro: 0.873)
```





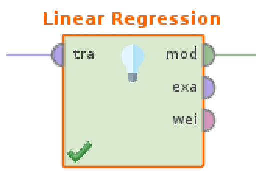
Residuals should be near-normally distributed

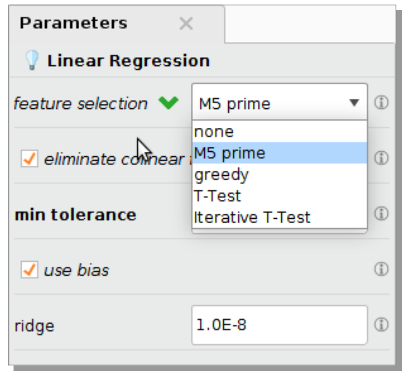
Also see KD 5.1.2 on Regression interpretation

DEAKIN BUSINESS SCHOOL

Selection of Attributes

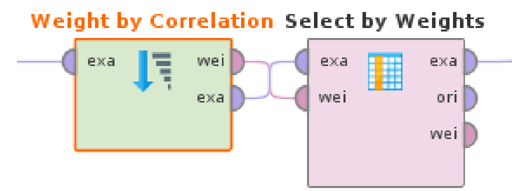
- Regression method require some way of selecting attributes for the model creation.





- Regression models commonly use a stepwise ("greedy") method of attributes selection, i.e. adding or removing attributes to improve the model performance.
- Greedy (or locally optimised) selection is not guaranteed to find the optimum solution.
- Stepwise selection or elimination of attributes usually also leads to model over-fitting the training data.
- An alternative to greedy attribute selection is "M5 prime", which is very effective and a default in RM. It uses regression trees to select the best attributes.

- Selection of the best attributes is called *feature engineering*, including:
  - Feature weighing
  - Feature selection/generation (optimised)



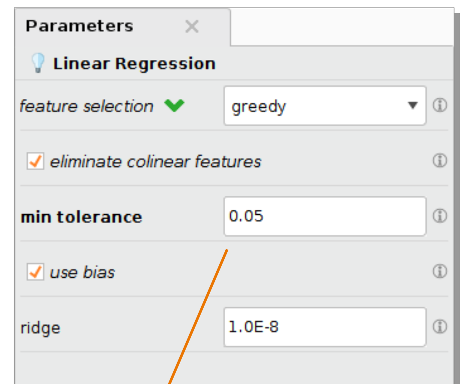
- The simplest is the *feature weighing*.
- In this approach, attributes are weighed against the label using correlation (high weight = high label and predictor correlation), then you can select top k attributes.
- One way to guarantee selection of the best attributes is to use *brute force*, i.e. trying all possible attrs combinations. This is computationally prohibitive.
- A better approach is to use *evolutionary feature engineering*, which aims to optimise the model by selecting or generating best features (e.g. using a genetic algorithm).
- However, *never discount the simple* way of selecting attributes –the simplest approach can sometimes give the best result (or close to optimal).

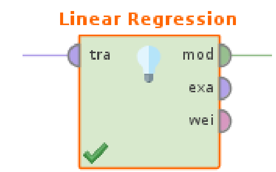
Multi-Collinearity: Tolerance, VIF and R^2

- One of the regression requirements is independence of predictor attributes.
- Analysis of pairwise correlation between predictors in not always sufficient for determining their independence.
- It is possible that a predictor is a linear combination of other predictors; in other words, it is possible to create a regression model to predict one predictor using the others, so that its *coefficient of determination* R^2 is high, i.e. $R^2 > 0.8$.
- The majority of systems calculate the following statistics for each predictor:

$$\textit{Tolerance} = (1 - R^2)$$
$$\textit{VIF} = 1 / (1 - R^2)$$
$$\textit{VIF} = 1 / \textit{Tolerance}$$

- R^2, tolerance and VIF (variance inflation factor) are clearly related.
- When for some attribute $R^2 > 0.8$ then *tolerance < 0.2* and *VIF > 5*, if lower threshold of tolerance is required, e.g. for *R2 > 0.95*, then *tolerance < 0.05* and *VIF > 20*. In all such cases the attribute would be considered multi-collinearly dependent on other predictors and should be removed.





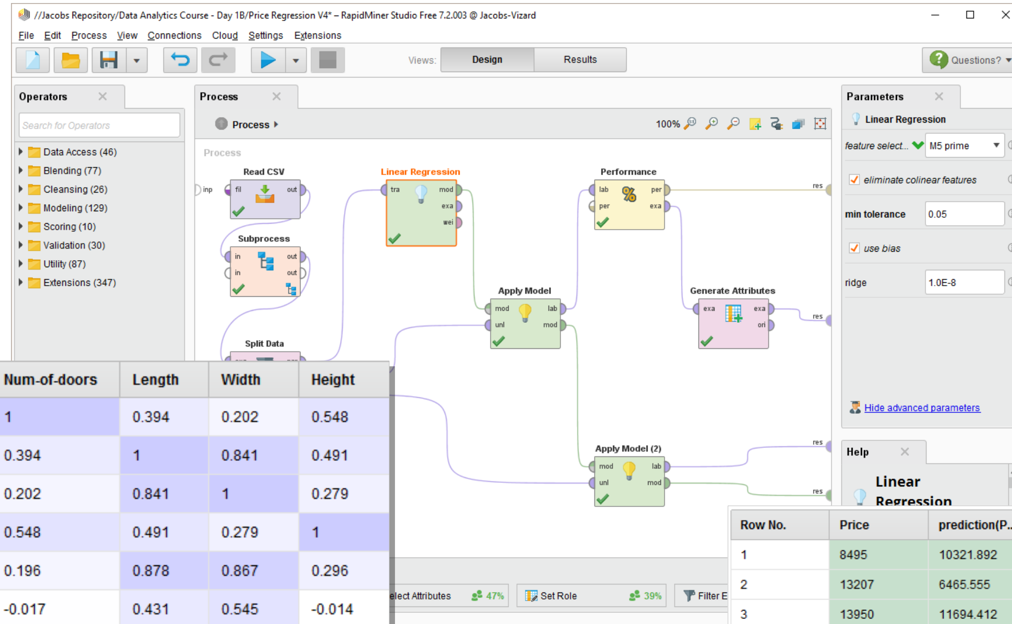
Attribute	Coefficient	Std. Error	Std. Coeffic...	Tolerance	t-Stat	p-Value	Code
MS_Zoning = RM	-7932.628	1627.039	-0.036	0.922	-4.875	0.000	****
Lot_Shape = Reg	-2174.858	1102.222	-0.013	0.912	-1.973	0.049	**
Neighborhood = NridgHt	7350.332	2578.598	0.022	0.797	2.851	0.004	***
Neighborhood = NoRidge	32651.604	3642.117	0.062	0.932	8.965	0	****
Exterior_1st = VinylSd	-2134.778	1354.327	-0.013	0.866	-1.576	0.115	
Mas_Vnr_Type = None	9669.770	1527.897	0.059	0.802	6.329	0.000	****
Mas_Vnr_Type = Stone	5738.905	2033.239	0.020	0.901	2.823	0.005	***
Exter_Qual = Gd	3524.525	1742.496	0.021	0.785	2.023	0.043	**
Exter_Qual = Ex	27810.095	3885.745	0.065	0.771	7.157	0.000	****
Foundation = CBlock	-2021.385	1291.248	-0.012	0.871	-1.565	0.118	

In RapidMiner default minimum tolerance is 0.05

It means that RapidMiner is VERY tolerant!

Practice: RapidMiner Studio

Process



Correlation

Attributes	Num-of-doors	Length	Width	Height
Num-of-doors	1	0.394	0.202	0.548
Length	0.394	1	0.841	0.491
Width	0.202	0.841	1	0.279
Height	0.548	0.491	0.279	1
Curb-weight	0.196	0.878	0.867	0.296
Num-of-cylinders	-0.017	0.431	0.545	-0.014
			0.735	0.067

Performance

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 3134.683 +/- 0.000
relative_error: 16.57% +/- 13.61%
root_relative_squared_error: 0.334
squared_correlation: 0.891
prediction_ave
```

LinearRegression

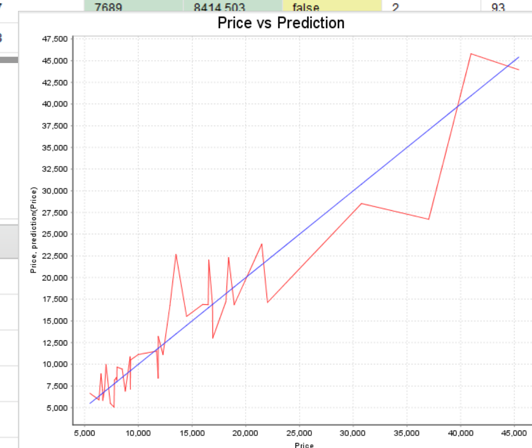
```
  94.836 * Wheel-base
+ 18.557 * Length
+ 86.390 * Width
+ 76.989 * Height
- 0.572 * Curb-weight
- 2164.052 * Num-of-cyli
+ 216.332 * Engine-size
- 6772.350 * Bore
- 5197.469 * Stroke
+ 277.381 * Compression-
+ 24.093 * Horsepower
+ 2.578 * Peak-rpm
+ 142.098 * City-mpg
- 271.941 * Highway-mpg
+ 77.187
```

Regression Formula

RapidMiner Studio

Prediction

Row No.	Price	prediction(P...	outlier	Num-of-doors	Wheel-base
1	8495	10321.892	false	4	98.800
2	13207	6465.555	false	2	94.500
3	13950	11694.412	false	4	99.800
4	15690	22076.142	false	4	104.500
5	8238	7113.387	false	2	94.500
6	5572	6653.980	false	2	93.700
7	7689	8414.503	false	2	93
8					



Diagnostic Charts

Attribute	Coefficient	Std. Error	Std. Coefficie...	Tolerance	t-Stat		
Wheel-base	94.836	140.348	0.078	0.676	0.676		
Length	18.557	77.275	0.032	0.454	0.240		
Width	86.390	341.804	0.026	0.396	0.253		
Height	76.989	186.831	0.027	0.961	0.412		
Curb-weight	-0.572	2.877	-0.040	0.177	-0.199		
Num-of-cylind...	-2164.052	910.790	-0.278	0.409	-2.376		
Engine-size	216.332	38.966	1.022	0.205	5.552	0.000	****
Bore	-6772.350	2240.012	-0.252	0.596	-3.023	0.003	***
Stroke	-5197.469	1201.488	-0.224	0.977	-4.326	0.000	****
Compression-...	277.381	112.347	0.159	0.999	2.469	0.015	**

Coefficients

DEAKIN BUSINESS SCHOOL

Summary / Review

- What are regression model assumptions / requirements?
- What is correlation?
  How different is correlation from causation?
- Why is Pearson correlation useful in regression analysis?
- Can Pearson correlation be applied to nominal attributes?
- Explain regression terms: intercept, slope and residuals.
- What is the difference between Pearson correlation (r) and coefficient of determination (R^2)?
- Explain regression pros and cons.
- How can you use scatter plot in regression analysis.
- What are extreme values?
  What is their impact on regression modelling?
- What is multi-collinearity?
  Is it the same as correlation?
  What needs to be done about it?
- What is tolerance?
  How is it used?
- What is dummy encoding?
  What should be done when we get too many dummy variables?
- Explain the role and method of coefficient analysis.
- Explain the main issues of regression model preparation in business contexts.

DEAKIN BUSINESS SCHOOL