Predictive Analytics – Session 3

Machine Learning Tools: Supervised Learning
Introduction to Unsupervised Learning

Associate Professor Ole Maneesoonthorn

Associate Professor in Econometrics and Statistics

Melbourne Business School

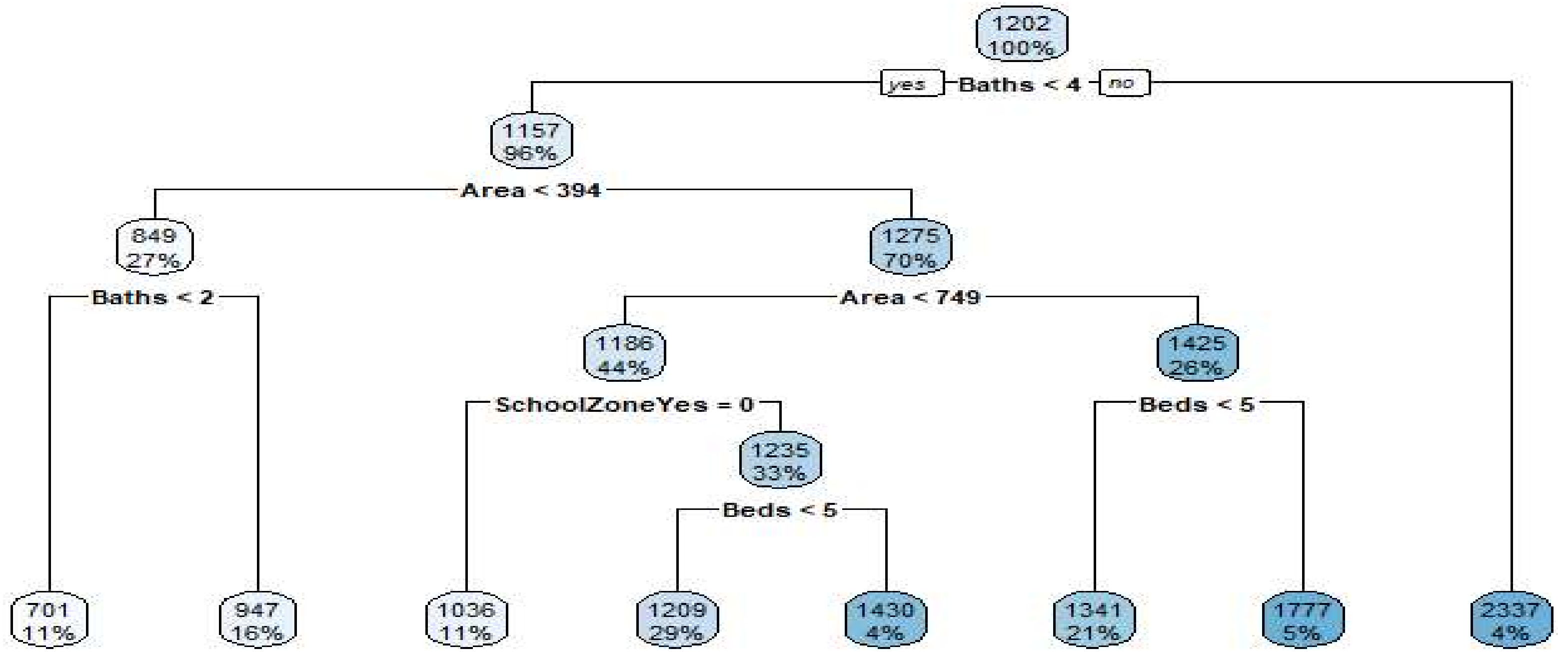
O.Maneesoonthorn@mbs.edu

Predicting Real Estate Prices with Machine Learning Tools

The Predictive Tools

Regression Trees

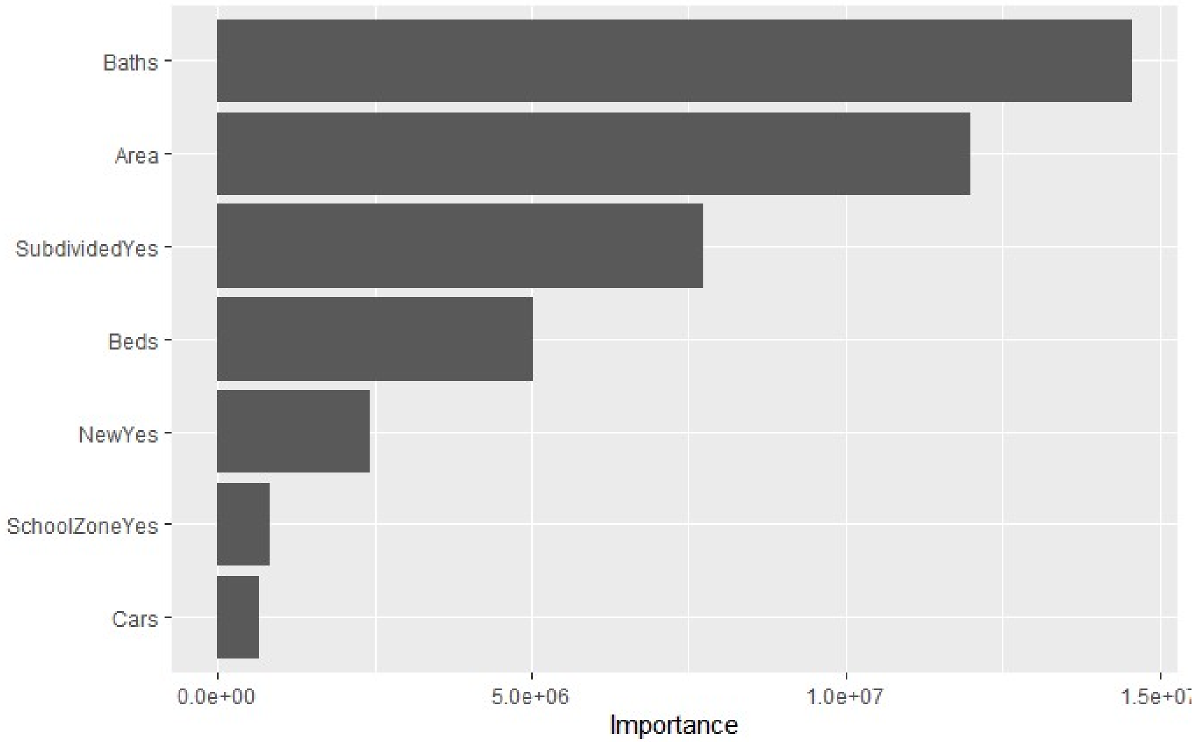
- Example: Predicting real estate prices



The Predictive Tools

Regression Trees

- Variable importance plot: high to low



The Predictive Tools

Regression Trees

If the tree is too large, consider **pruning**

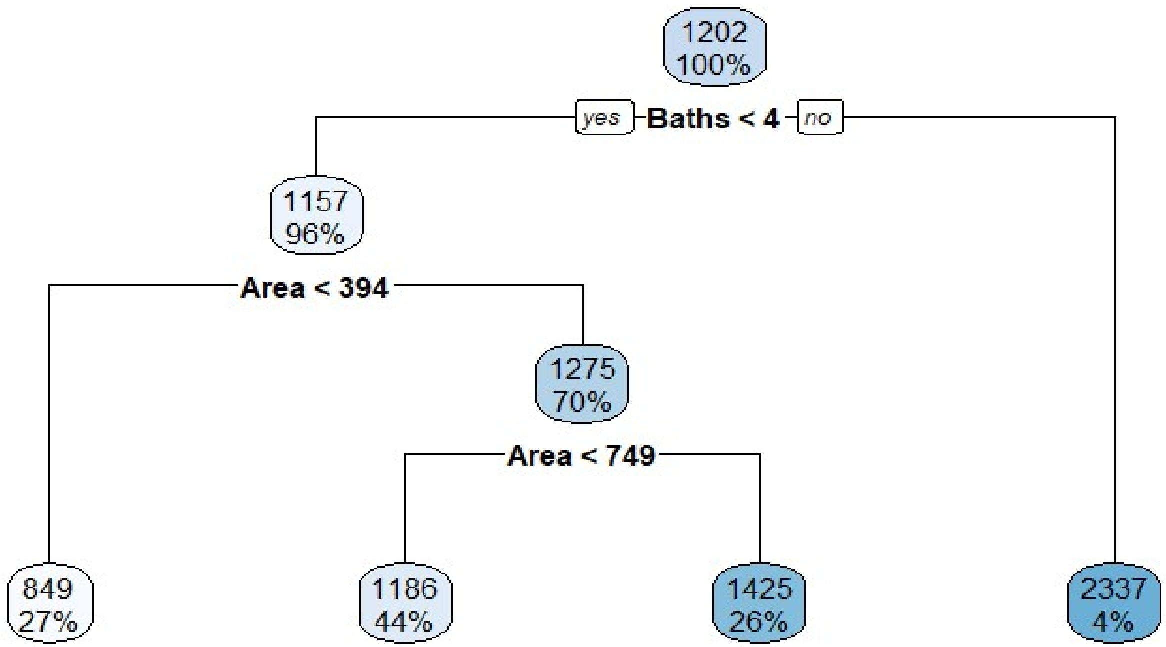
- That is, cut of some unnecessary branches

- Prune by defining a larger value of **complexity parameter (CP)**
    - The default in R is CP=0.01
    - Again, the **larger CP, the smaller your pruned tree**

The Predictive Tools

Regression Trees

- Example: Predicting real estate prices, tree pruned with CP=0.05



The Predictive Tools

Regression Trees

- Predicting with regression trees:

Follow the relevant branches!

- Nature of the relationship depends on the construct of the tree – can be difficult to work out

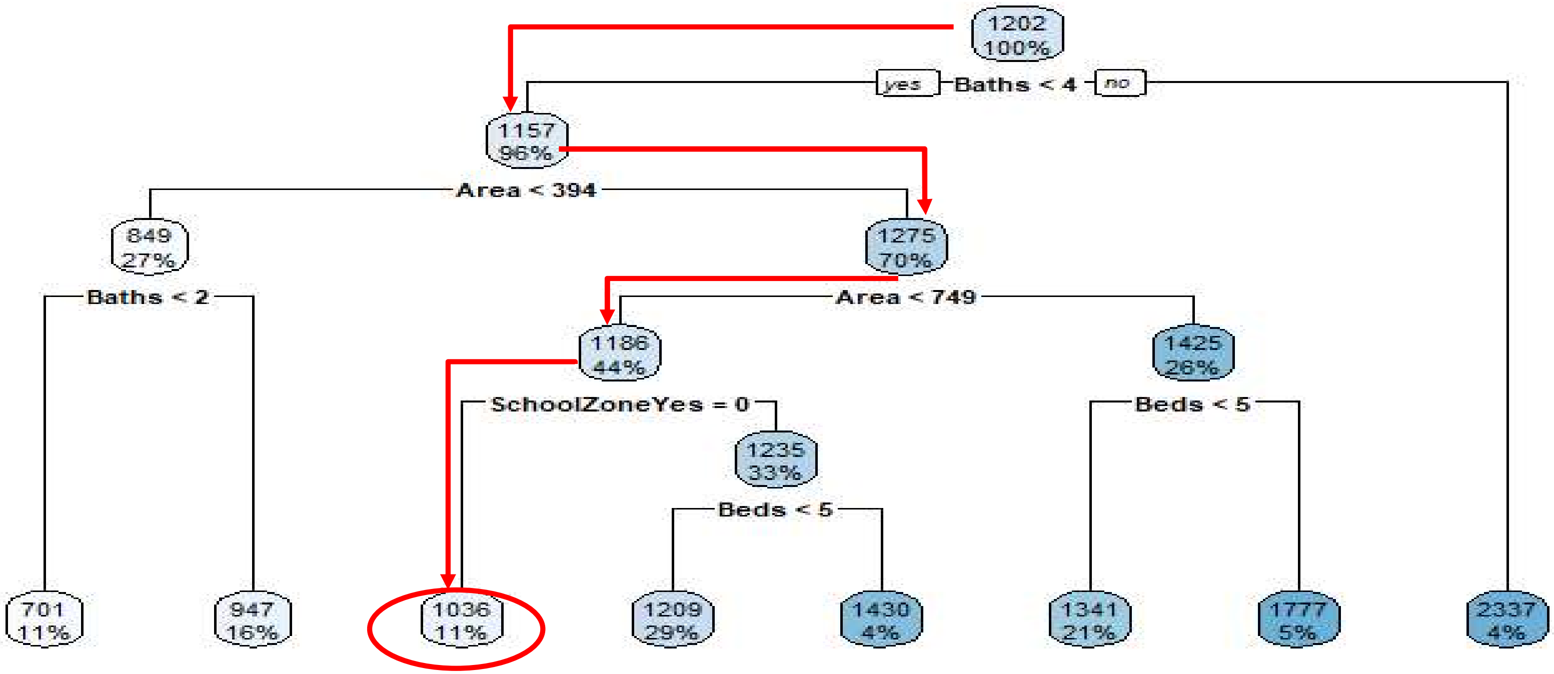
The Predictive Tools

Regression Trees

- Example: Predicting real estate prices

Let's predict the sales price of a property with:

- 720sqm area, not subdivided
- 3 bathrooms, 5 bedrooms, 2 car spaces
- Private sale
- Not in school zone



The Predictive Tools

Neural Network

- Example: Predicting real estate prices (Real Estate Data)
- Let us construct a neural network with one hidden layer
- Allow for 4 hidden neurons (nodes)
- Inputs variables and output variable are scaled using z-score, except for dummy variables

The Predictive Tools

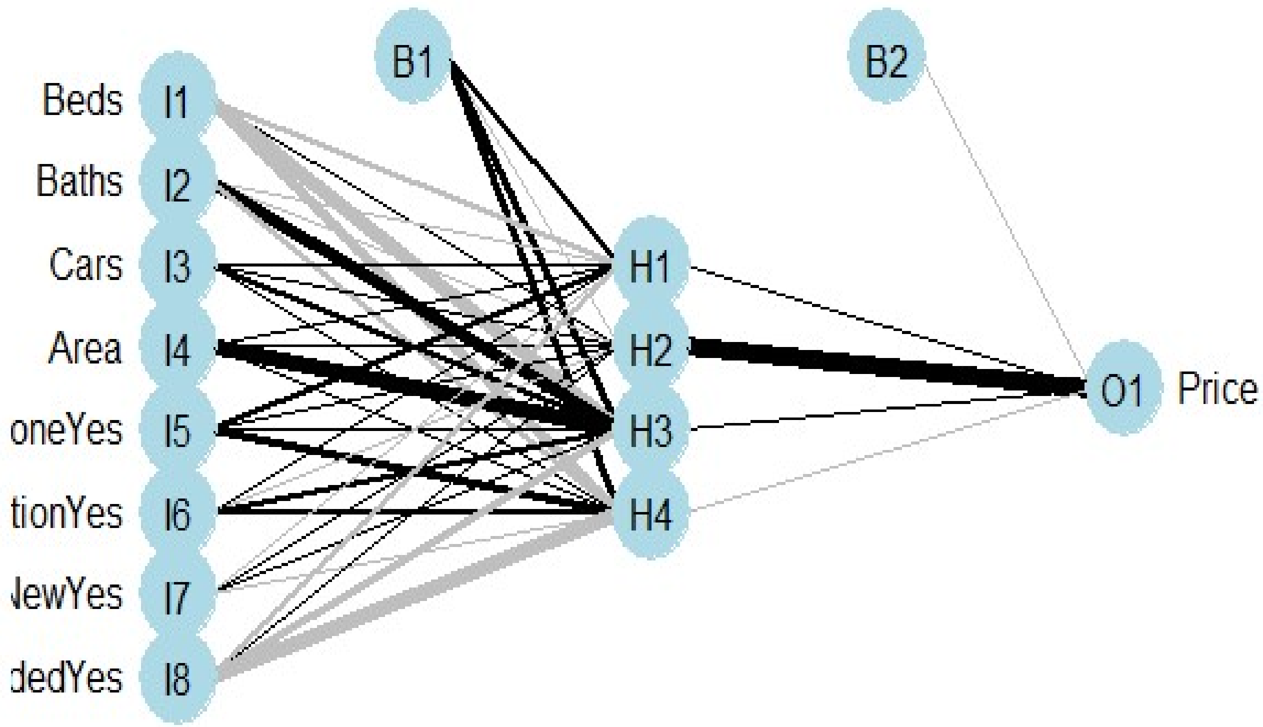
Neural Network

- The training network
  - Grey = negative weight
  - Black = positive weight
  - Thickness → magnitude

- Does this mean much??

- Prediction relies on the black-box computation behind the scene

Network Plot



The Predictive Tools

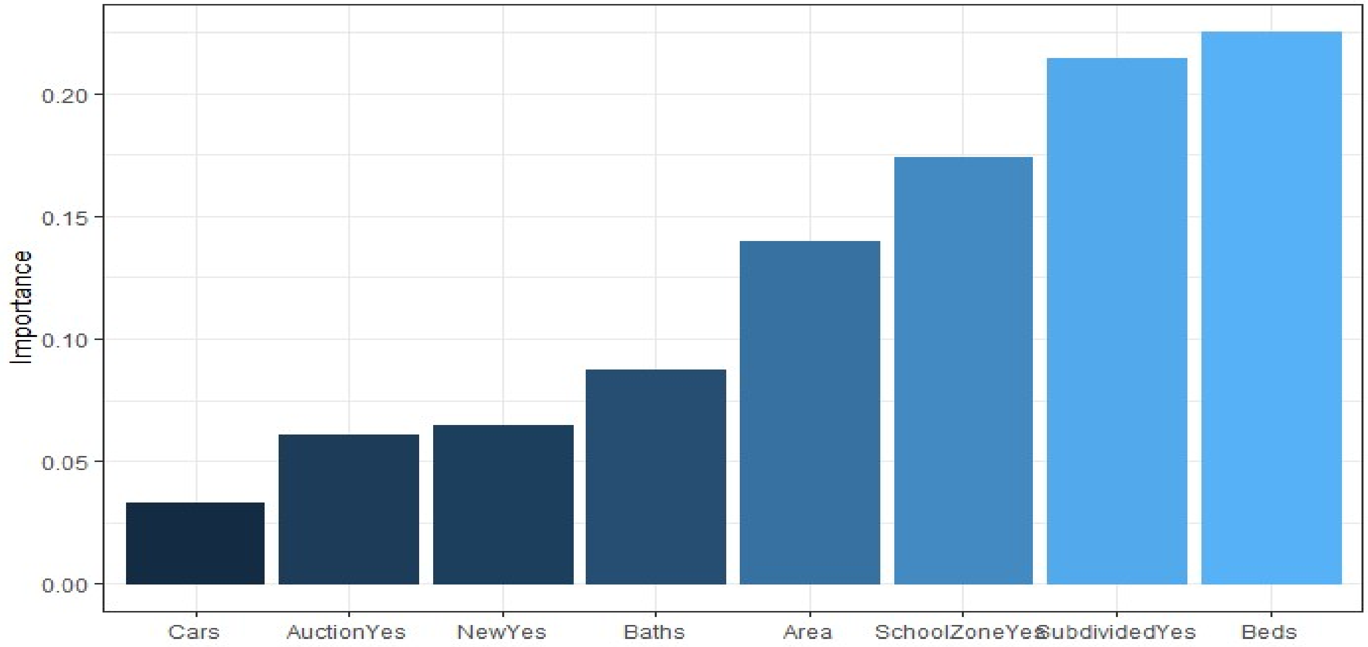
Neural Networks

- **Garson's relative importance**
  - Measure the contribution in absolute relative magnitude
  - Ranges from 0 to 1 – higher indicate more important
  - Only applicable to network with one hidden layer and one output node
  - Direction of response cannot be determined

The Predictive Tools

Neural Networks

- Garson's relative importance
  - Top four importance: beds, subdivision, school zone & area
  - Least important: car spaces & auction



The Predictive Tools

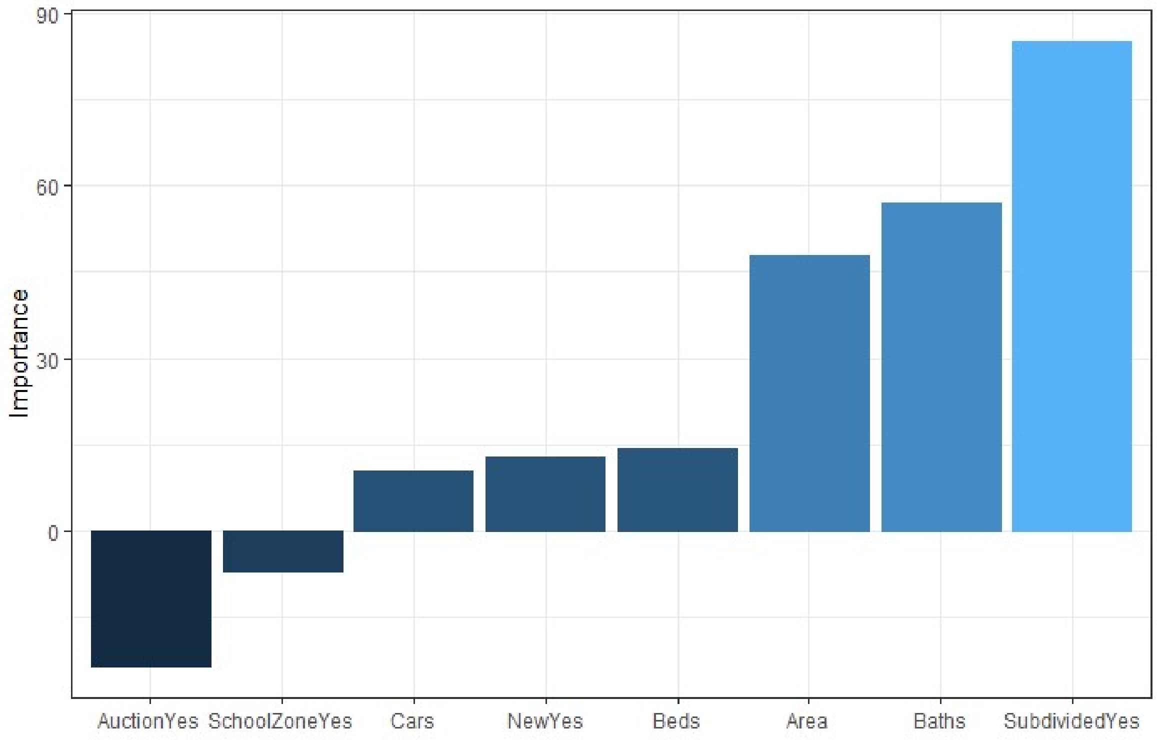
Neural Networks

- **Olden's connection weights**
  - Sum of the product of input-hidden and hidden-output weights
  - Unit is sensitive to data scale and linking functions
  - Positive & negative weights will cancel
  - Can handle multiple hidden layers

The Predictive Tools

Neural Networks

- Olden's connection weights
  - School zone has an overall negative impact (??)
  - Subdivision premium
  - Area and number of bathrooms have positive impact
  - Remember that all data are in z-score scale



The Predictive Tools

Neural Networks

- **Sensitivity analysis – Lek profile**
  - Set all explanatory variables fixed
  - Sequence the variable to interest from min to max
  - Predict the outcome
  - Plot to look into how the outcome changes as the variable of interest goes from min to max

The Predictive Tools

Neural Networks

Lek profile – how do we fix the ‘other’ variables?

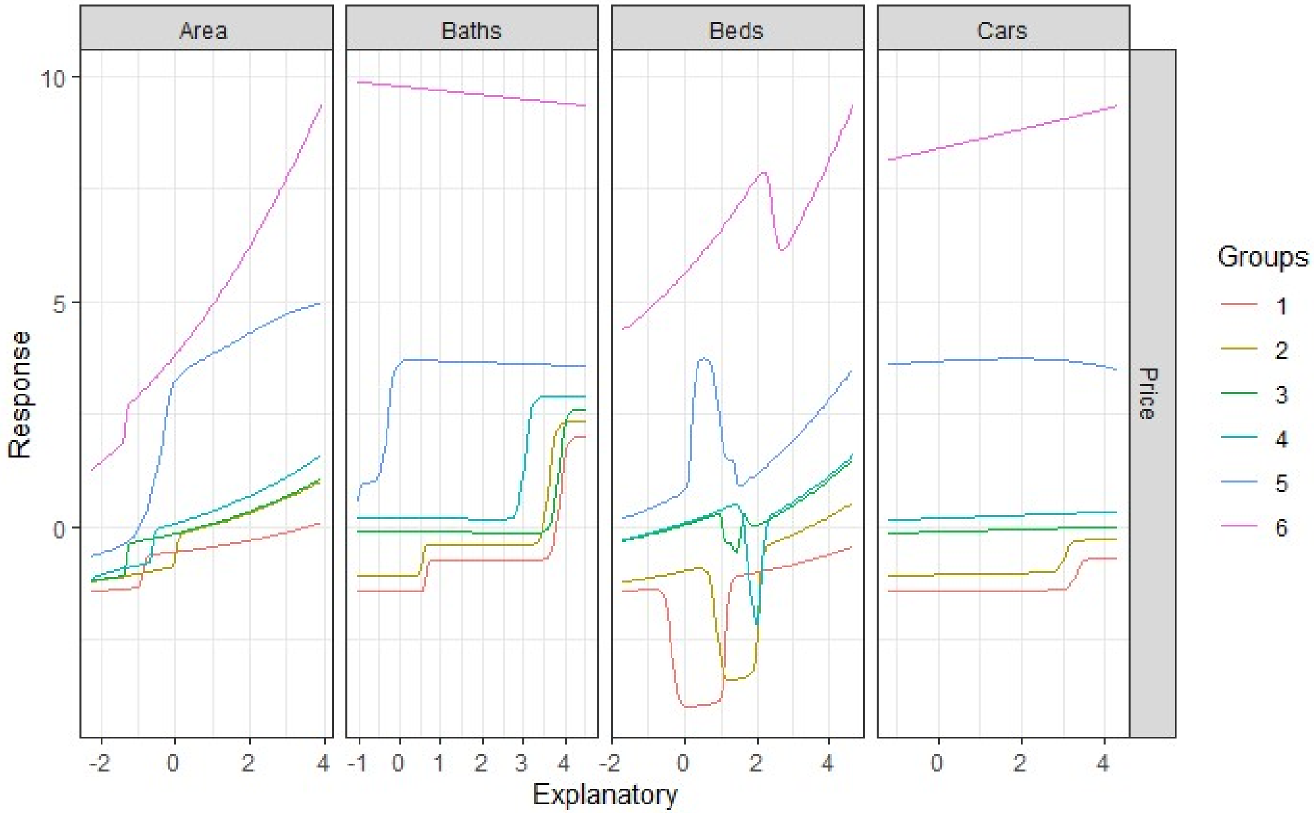
1. Group by quantiles – fix at some percentiles
   – Default: min, 20^{th}, 40^{th}, 60^{th}, 80^{th}, max
   – You can change these

2. Group by clusters
   – If variables are related in a complex manner, grouping by quantiles may not make sense
   – E.g. when X1 is at the max, X2 may tend to be in its lower range
   – Group by cluster uses k-means clustering and fix the ‘other’ variables at the centre of each cluster
   – You just need to determine and number of clusters to use

The Predictive Tools

Neural Networks

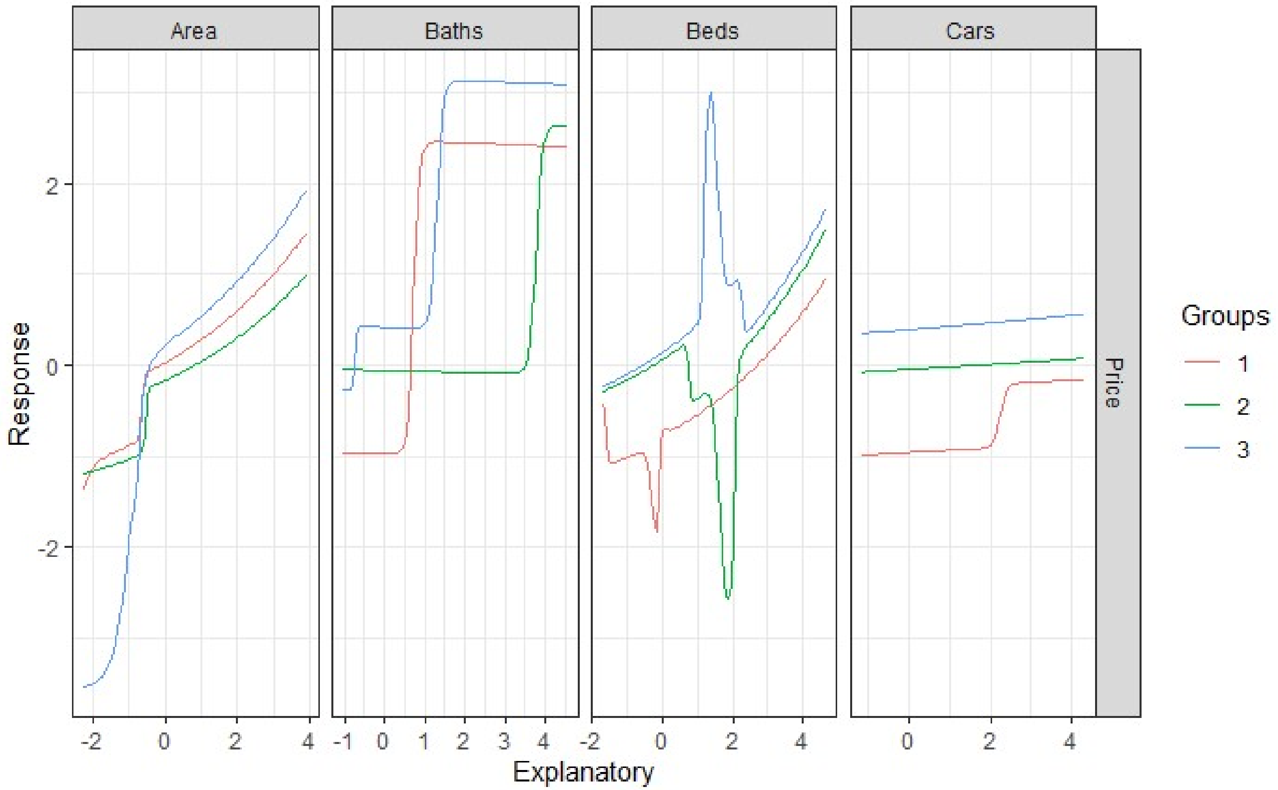
- Lek profile – default quantile grouping



The Predictive Tools

Neural Networks

- Lek profile – grouping by clusters (3 clusters)



The Predictive Tools

Neural Networks

The real test will be the predictive accuracy assessment!

The Predictive Tools

The Real Estate Price

Predictive performance so far…

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	**75.437**
Stepwise	295.628	186.304	13.940	0.593	260.297	77.586
Nonlinear	**280.542**	**175.747**	**13.387**	**0.560**	**245.350**	80.461
RegTree	305.524	212.402	16.759	0.676	306.196	109.889
NNet	393.638	247.676	18.430	0.789	353.798	125.998

Green is best

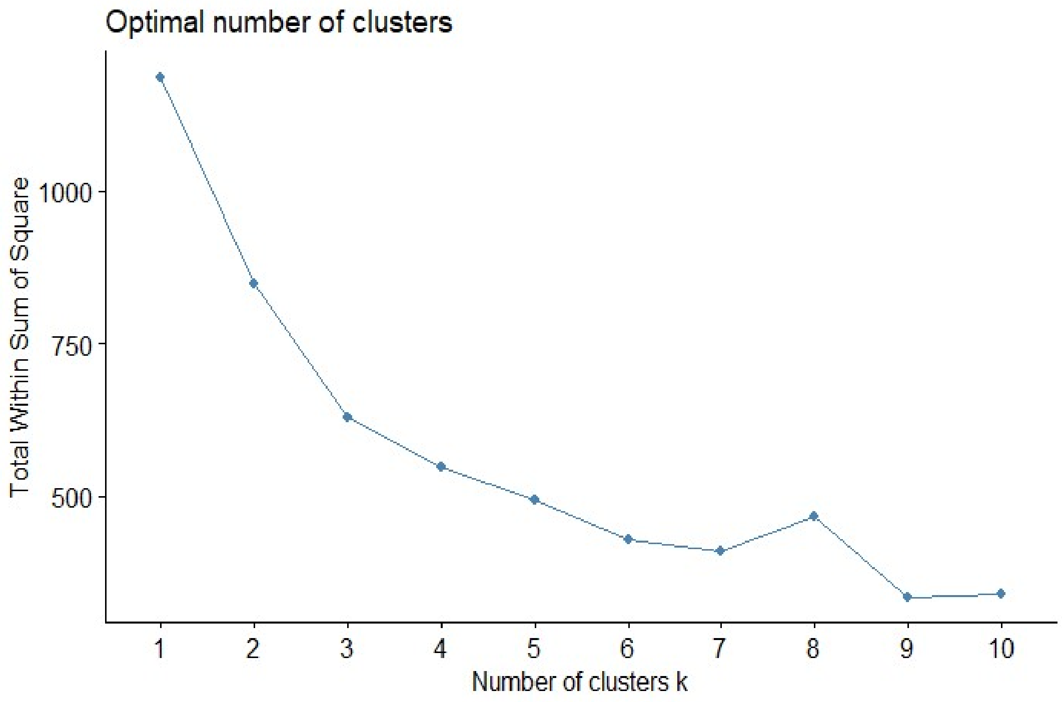
Red is worst

The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: We first have to decide the value of K

- **The Elbow method**
  - Generally, larger K, smaller total WCSS
  - Select K when you see the elbow bend
  - i.e. when the slope begins to flatten
  - Not so clear-cut in this data
  - My choice: K=6

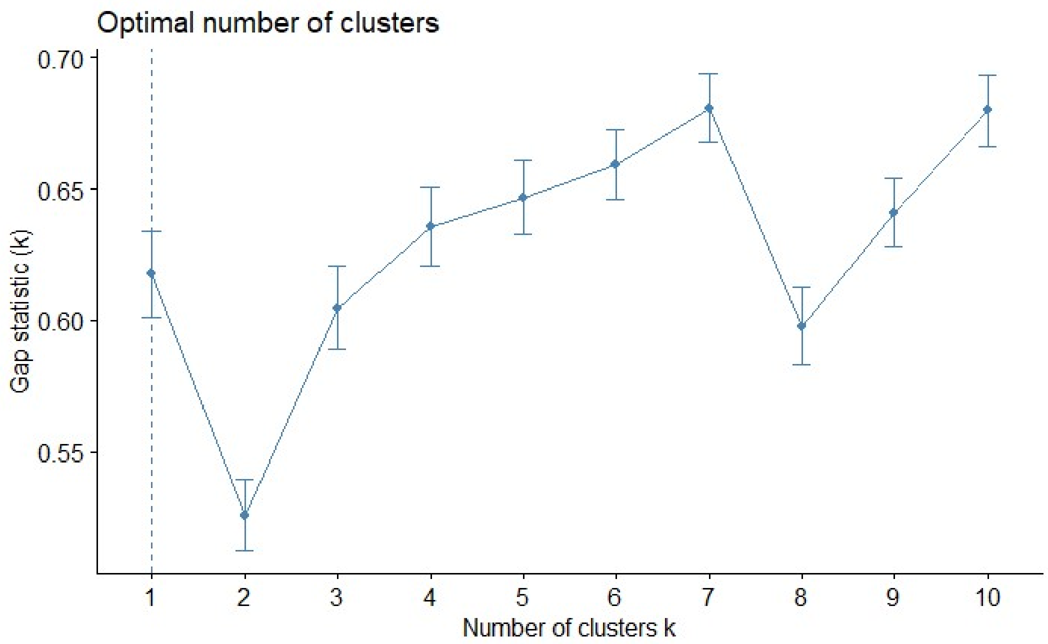


The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: We first have to decide the value of K

- **The Gap statistic** - The "gap" between k-means clusters and uniform (random) partition over the space
  - IDEAL: maximum Gap statistic
  - REALITY: there are variations in data
  - Look at Std. Error of statistic as well
  - Choose smallest "K" such that $Gap(K) \geq Gap(K+1) - Std.Err.$

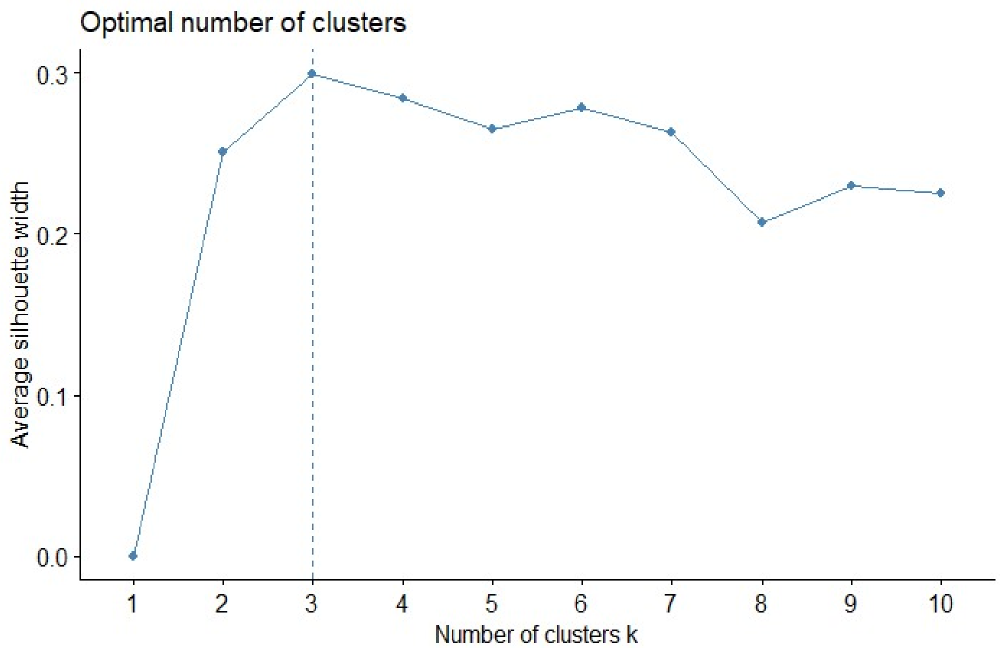


The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: We first have to decide the value of K

- **The Silhoutte measure** – measure how similar an observation is to the other observations in the same cluster
- Choose "K" with the highest value!

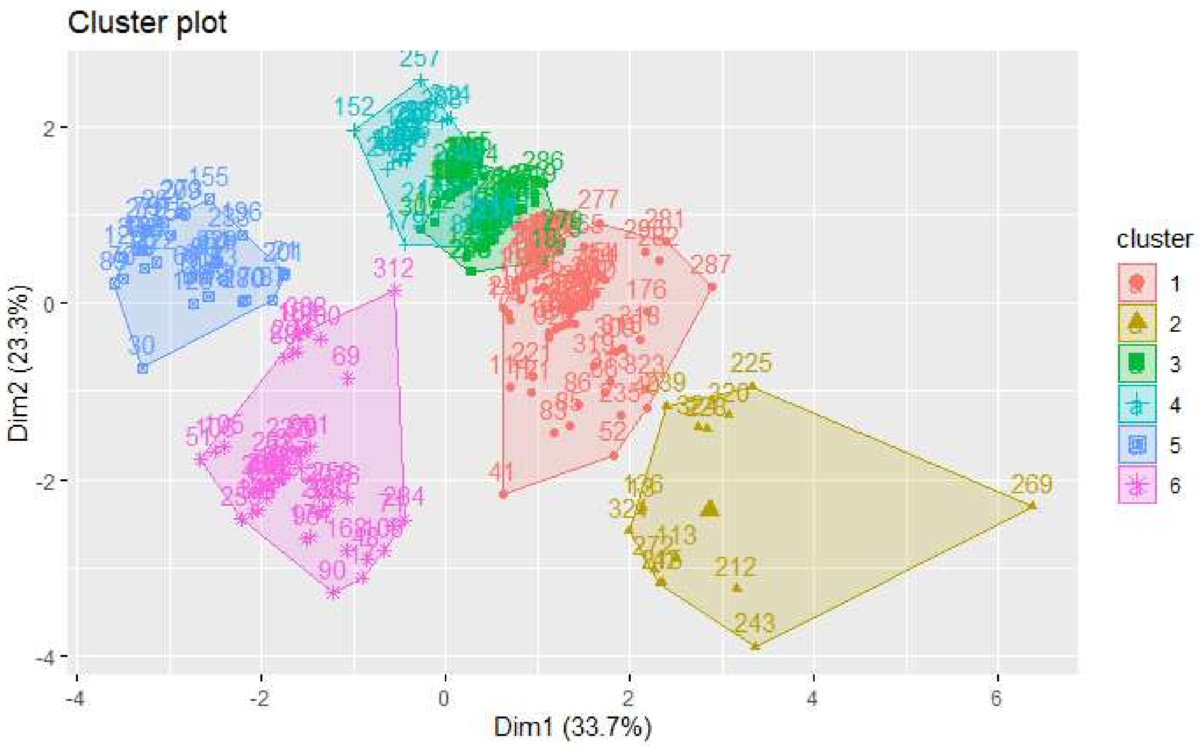


The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: K-means with 6 clusters (based on Elbow method)

- Visualisation of the clusters



The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: K-means with 6 clusters

- Cluster characteristics

cluster	Price	Beds	Baths	Cars	Area	SchoolZoneYes	AuctionYes	NewYes	SubdividedYes
1	1350.64	4.34	2.33	2.04	754.33	0.83	0.49	0.13	0.00
2	2124.60	5.13	4.07	3.00	780.13	0.80	0.47	0.60	0.00
3	1186.74	3.15	1.32	2.11	746.04	0.81	0.47	0.00	0.00
4	1217.31	3.18	1.49	1.00	741.69	0.87	0.46	0.00	0.00
5	723.12	2.39	1.09	1.18	284.73	0.73	0.55	0.03	1.00
6	1007.57	3.40	2.15	1.87	327.53	0.77	0.66	0.81	0.98

The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: K-means with 6 clusters

- Test set prediction with **Option 1**
- → There will only be six unique predicted values!
- E.g. Test set observation ID #3 belongs to cluster 4
  - Its predicted price is $1.217m
- Test observation ID #28 also belongs to cluster 4
  - Its predicted price is also $1.217m
- Even though the two properties have different characteristics

cluster	Price
1	1350.64
2	2124.60
3	1186.74
4	1217.31
5	723.12
6	1007.57

The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: K-means Prediction **Option 2**

- Let us fit the regression model to each cluster
- But we will consider reducing the number of clusters for this option
  - It turns out that using K=2 is best under this option
- Note: We need to remove the variable that “perfectly” defines a cluster
  - To avoid multicollinearity in the regression model
  - For K=2 we remove the SubdivisionYes variable

The Predictive Tools

Unsupervised Learning – K-means

Real Estate Price Example: K-means Predictive Accuracies

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	75.437
Stepwise	295.628	186.304	13.940	0.593	260.297	77.586
Nonlinear	280.542	175.747	13.387	0.560	245.350	80.461
RegTree	305.524	212.402	16.759	0.676	306.196	109.889
NNet	393.638	247.676	18.430	0.789	353.798	125.998
kMeans	357.701	236.212	18.424	0.752	348.141	121.583
kMeans(Reg)	285.684	182.035	13.637	0.580	253.972	75.955

Green is best

Red is worst

The Predictive Tools

Unsupervised Learning – k-Nearest Neighbour (k-NN)

Real Estate Example

- As in the k-means, we conduct k-nn regression on scaled data
  - The issue of multiple units of measurements & collective distance measures

- We apply k-nn regression with $K = 16$

- Training set predictability → Predictive $R^2 = 0.552$
  - Recall: lnear regression R^2 approx. 0.64 on the training set

The Predictive Tools

Unsupervised Learning – k-Nearest Neighbour (k-NN)

Real Estate Example - Predictive performance?

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	**75.437**
Stepwise	295.628	186.304	13.940	0.593	260.297	77.586
Nonlinear	**280.542**	**175.747**	**13.387**	**0.560**	**245.350**	80.461
RegTree	305.524	212.402	16.759	0.676	306.196	109.889
NNet	393.638	247.676	18.430	0.789	353.798	125.998
kMeans	357.701	236.212	18.424	0.752	348.141	121.583
kMeans(Reg)	285.684	182.035	13.637	0.580	253.972	75.955
K-nn	327.934	206.007	15.429	0.656	281.265	98.401

Green is best

Red is worst