Predictive Analytics – Session 4

Machine Learning Tools: Unsupervised Learning
Ensembles Methods

Associate Professor Ole Maneesoonthorn

Associate Professor in Econometrics and Statistics

Melbourne Business School

O.Maneesoonthorn@mbs.edu

Predicting Real Estate Prices with Ensembles and SVM

The Predictive Tools

Ensembles Methods – Bagging

Real Estate Prices

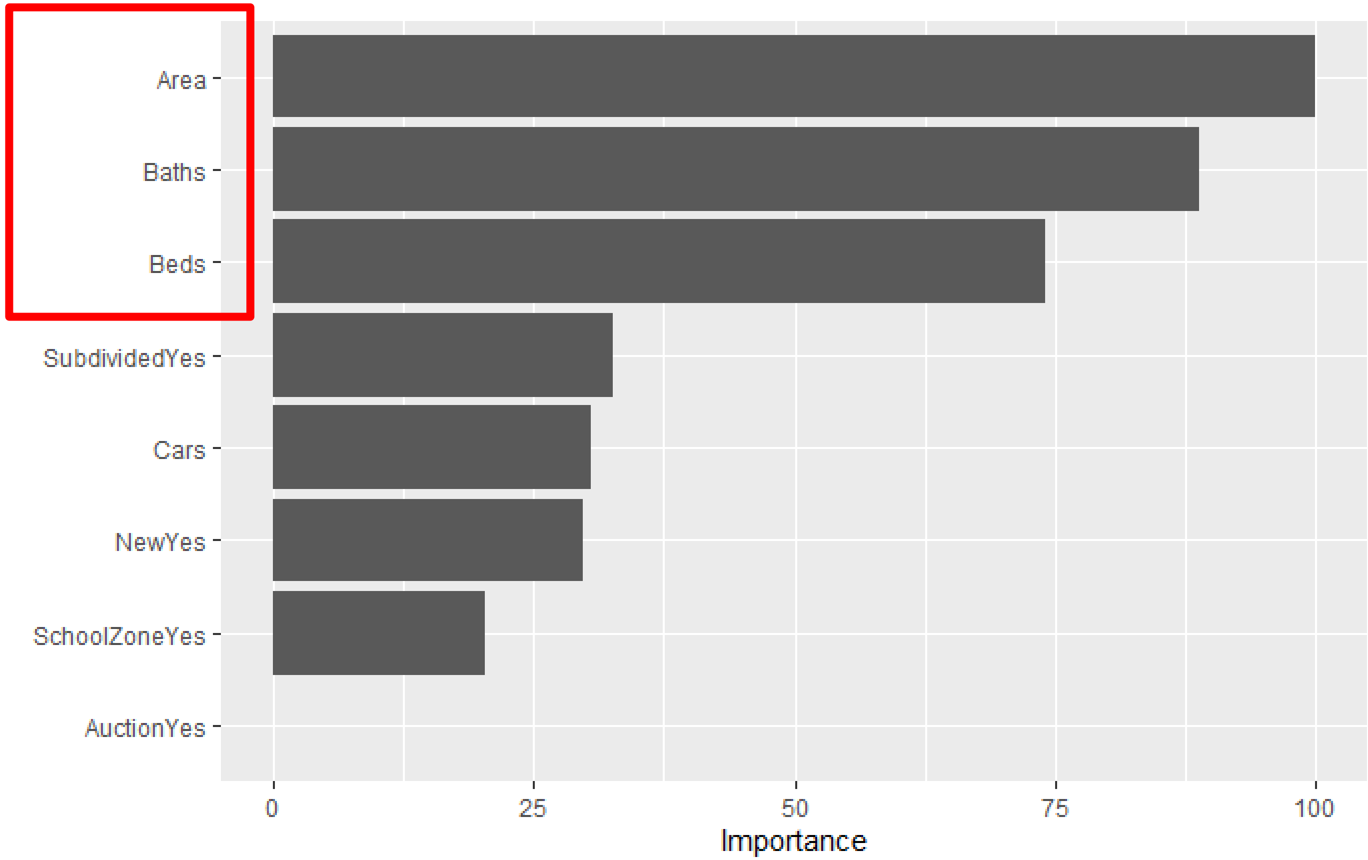
- Trained using “caret” library
- Base model: regression tree
- With 25 resamples → predictive $R^2 = 61.8\%$

The Predictive Tools

Ensembles Methods – Bagging

Real Estate Prices

- Variable importance
- Top 3: Area, Baths and Beds



The Predictive Tools

Ensembles Methods – Bagging

Real Estate Prices

- Predictive performance relative to regression tree

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
RegTree	305.524	212.402	16.759	0.676	306.196	109.889
Bagging(Tree)	286.877	184.752	14.035	0.588	258.901	87.470

- Bagging is significantly better at test set predictive accuracy!

- Full comparison to come…

The Predictive Tools

Ensembles Methods – Bagging

Pros

- Reduced variance means more stable predictions
- Can avoid over-fitting with a single model

Cons

- Lack of interpretation
- All base models are of the same class
- No guarantee that all training data points will be used

The Predictive Tools

Ensembles Methods – Random Forest

Real Estate Prices

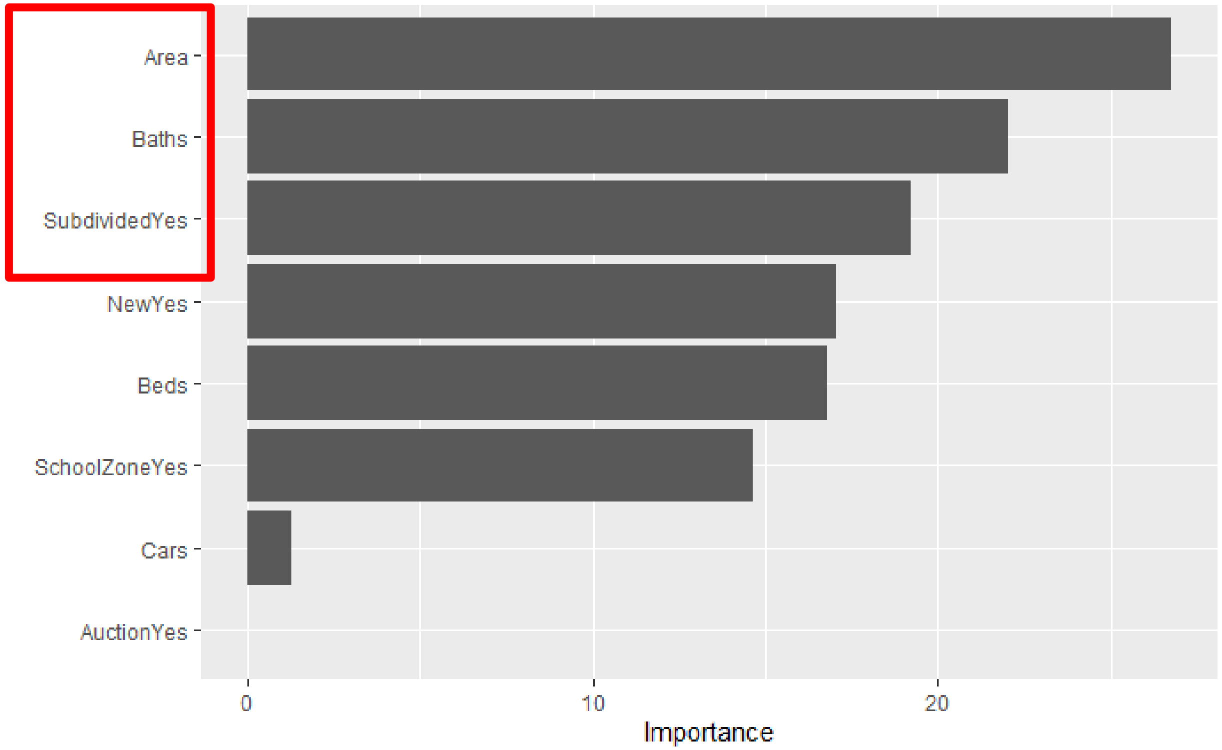
- Trained using “randomForest” library
- Need to specify the number of randomly selected input variables to add to the end of each tree – “mtry”
  - This has to be less than the number of input variables in the training set
  - Can be larger for cases where you have more inputs
  - This is an additional random component relative to bagging tree

The Predictive Tools

Ensembles Methods – Random Forest

Real Estate Prices

- % Variation explained = 64.1%
    - This is predictive R^2
- Variable importance
- Top 3: Area, Baths and Subdivided



The Predictive Tools

Ensembles Methods – Bagging

Real Estate Prices

- Predictive performance relative to regression tree

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
RegTree	305.524	212.402	16.759	0.676	306.196	109.889
Bagging(Tree)	286.877	184.752	14.035	0.588	258.901	87.470
Random Forest	289.917	184.607	13.890	0.588	265.139	80.995

- Mixed results compared to bagging!
  → Pick according to your choice of loss

The Predictive Tools

Ensembles Methods – Random Forest

Pros & Cons

- See **pros** & **cons** of Bagging!
- **Additional Con**: algorithm involve a greater degree of randomness

The Predictive Tools

Ensembles Methods – Boosting

Real Estate Prices

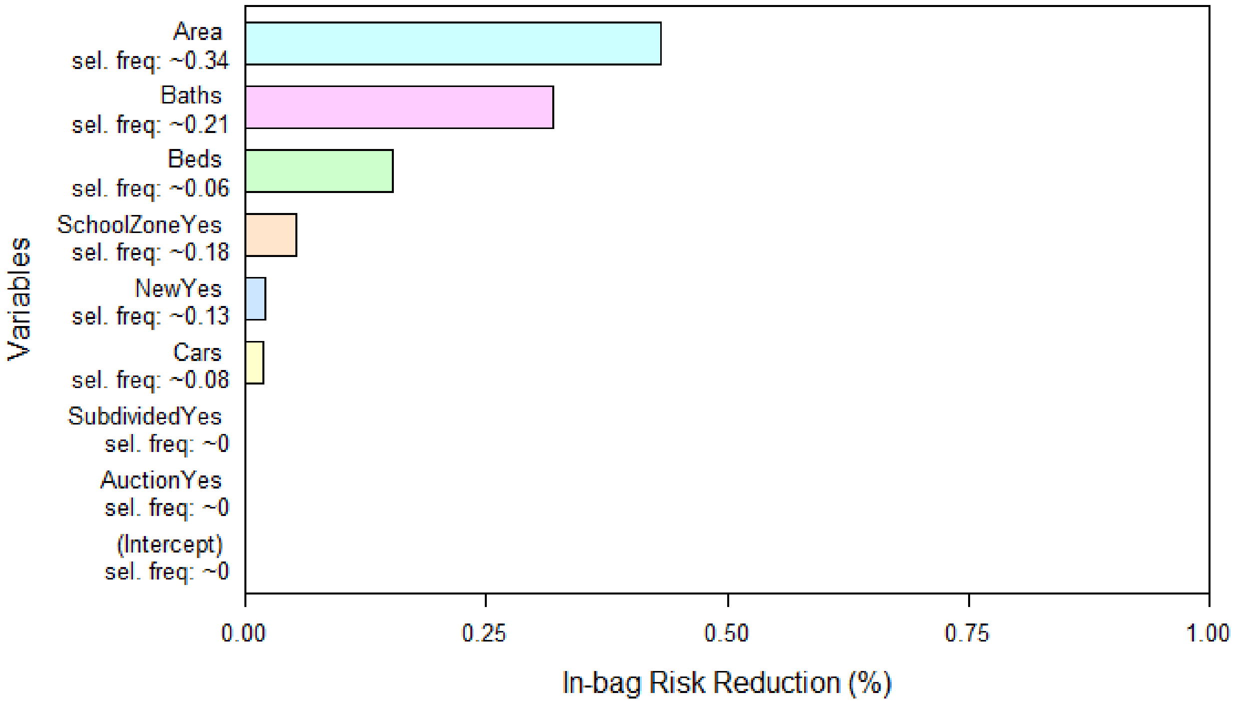
- The glmboost() function does component-wise boosting
- That is, there are “P” base models for “P” input variables
- At each boosting step the algorithm chooses to use a certain input if it reduces the “loss”
- → Not every variable is useful!

The Predictive Tools

Ensembles Methods – Boosting

Real Estate Prices

- Variable importance by risk reduction → how much does the input help to reduce loss?
- Note: frequently selected is not always coinciding with greater reduction of loss



The Predictive Tools

Ensembles Methods – Boosting

Real Estate Prices

- Predictive accuracy – compared to the linear regression

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	75.437
Boosting	307.306	181.584	13.291	0.578	251.016	70.219

- Boosting performs better in all metrics but the RMSE

The Predictive Tools

Ensembles Methods – Boosting

Pros

- You can work with simpler “weak learner” models
- Boosts predictive power of low predictability problems

Cons

- Limited interpretation – can only interpret the “base” learner
- Is boosting a “weak learner” really the final option? Or do we need to revisit the problem and data?

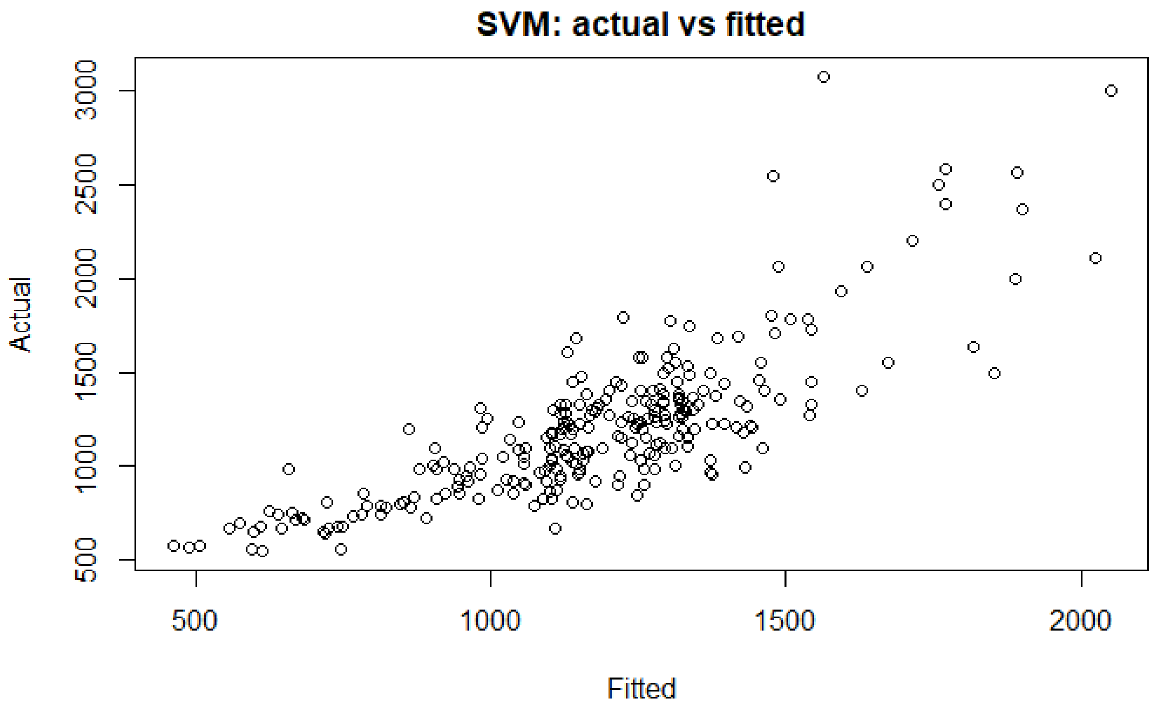
The Predictive Tools

Support Vector Machine (SVM)

Real Estate Prices

- A multiple input SVM model – let us look into actual vs fitted values

- Larger variation at higher prices



The Predictive Tools

Support Vector Machine (SVM)

Real Estate Prices

- Predictive accuracy – compared to the linear regression & boosting

	RMSE	MAE	MAPE	MASE	Banker	RE Agent
Linear	296.337	185.768	13.885	0.592	259.730	75.437
Boosting	307.306	181.584	13.291	0.578	251.016	70.219
SVM	303.605	177.032	12.820	0.564	230.667	68.925

- Boosting performs better in all metrics but the RMSE

The Predictive Tools

Support Vector Machine (SVM)

Pros

- Flexible predictive function
- Allow for discontinuity in the function via, e.g., radial kernels
- Usually very competitive in terms of predictability

Cons

- Lack of interpretation
- Issue of overfitting
- Predictions outside the range of the input data may be highly inaccurate

Real Estate Example – A comprehensive predictive comparison!

	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
Bagging(Tree)	286.877	184.752	14.035	0.588	258.901	87.470
Random Forest	289.917	184.607	13.890	0.588	265.139	80.995
Boosting	307.306	181.584	13.291	0.578	251.016	70.219
SVM	303.605	177.032	**12.820**	0.564	**230.667**	**68.925**

Green is best

Red is worst

The Predictive Tools

Key takeaways

Having a good understanding of your tools means you will be an effective user of those tools.

The Predictive Tools

Key takeaways

There is no general rule on choosing a predictive model

– focus on the suitability of the tools to your problem.

The Predictive Tools

Key takeaways

There is (almost always) a trade-off between model's flexibility and interpretability.