

Practical Predictive Analytics Seminar

Talex Diede

Session 4: Machine Learning Topics

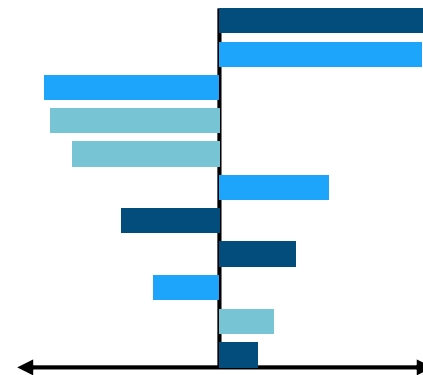
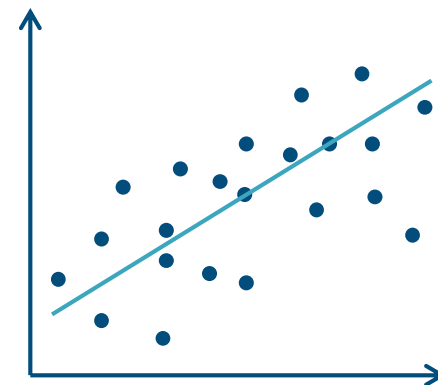
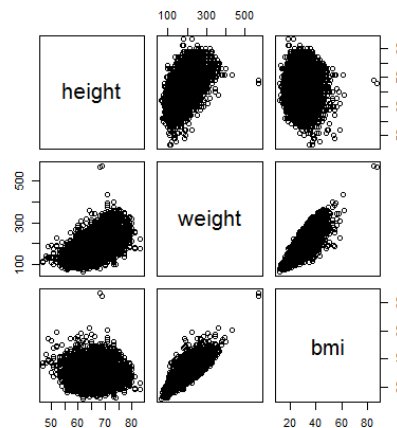
May 22, 2019



**SOCIETY OF
ACTUARIES**

GLM review

- Linear model
- Interpretable
- Issues:
 - Multicollinearity
 - Variable selection
 - Variable importance
 - Interactions



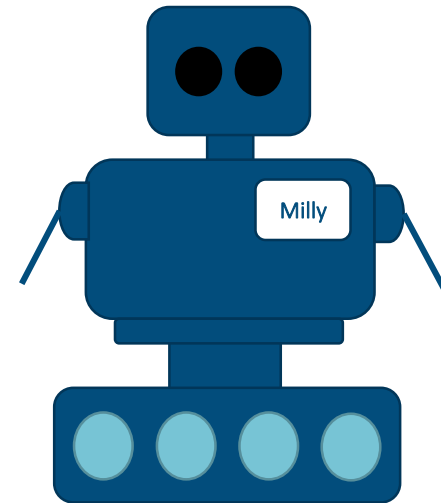
Why machine learning?

- Data continues to grow
- Powerful
- Flexible
- Computational enhancements
 - Cheaper
 - More available
- It's sexy



Machine learning techniques

- Regularization methods
- Classification and regression trees
- Ensemble models
- Others:
 - Clustering
 - Bayesian
 - Neural network
 - Deep learning

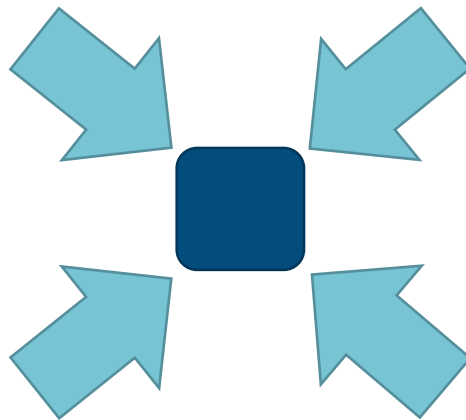


Regularization Methods



What is “regularization”?

- Regularization is a technique used to avoid the problem of overfitting. The idea is to add a complexity term to the loss function to penalize more complex models.



Regularization methods

- Ridge regression
- LASSO
- ElasticNet



- In R:
 - Packages: **glmnet**, MASS, ridge, lars, elasticnet, ...



Ridge regression



- weight decay
- L2-norm penalty

• *loglikelihood* =

$$-\sum [Y_i \ln(\hat{y}_i) + (1 - Y_i) \ln(1 - \hat{y}_i)] + \lambda \sum \beta^2$$



LASSO

- Least absolute shrinkage and selection operator
- L1-norm penalty
- *loglikelihood* =

$$-\sum [Y_i \ln(\hat{y}_i) + (1 - Y_i) \ln(1 - \hat{y}_i)] + \lambda \sum |\beta|$$



ElasticNet

- Convex combination of ridge and LASSO
- L2 & L1-norm penalties

- *loglikelihood* =

$$-\sum [Y_i \ln(\hat{y}_i) + (1 - Y_i) \ln(1 - \hat{y}_i)] + \lambda \left((1 - \alpha) \sum \beta^2 + \alpha \sum |\beta| \right)$$

Aside: Cross-Validation

- Useful for smaller datasets

1	2	3
4	5	6

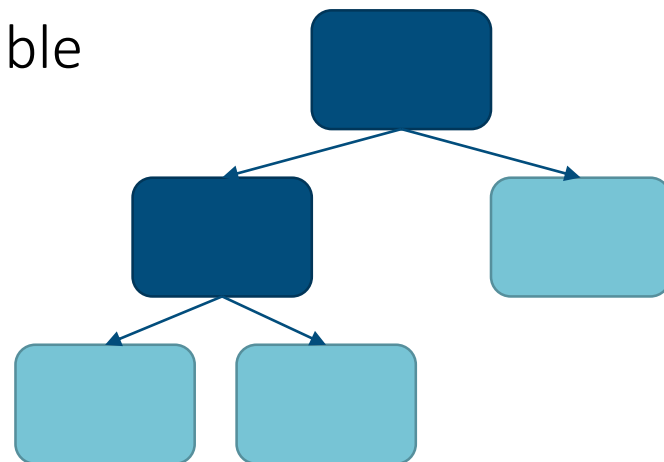
Classification and Regression Trees (CART)



Trees

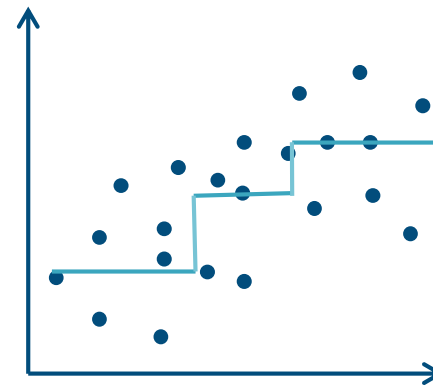
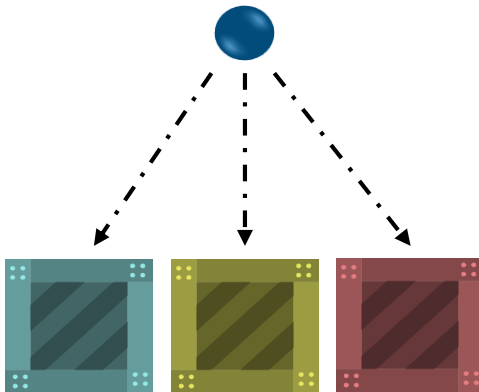
- Sequence of questions/rules for splitting the data
- Elements of CART algorithms
 - Rules for splitting data at each node
 - Stopping criteria
 - Prediction for the target variable

$N = 350$
 $0 = 200/350$
 $1 = 150/350$



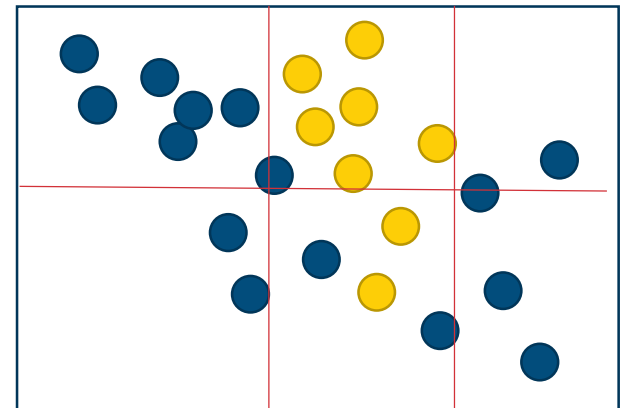
Classification vs regression

- Classification trees: used for categorical or binary target variables
 - Predict the category a policy will fall into
- Regression trees: continuous target variable
 - Predict the value of the continuous target



Splitting nodes

- Goal: choose the split that results in nodes with maximum homogeneity
- Classification: “Impurity” function
 - Entropy
 - Misclassification rate
 - Gini index
 - Twoing
- Regression: Squared residuals minimization



Stopping rules

- Depth
- Size
- Number of nodes
- Complexity parameter

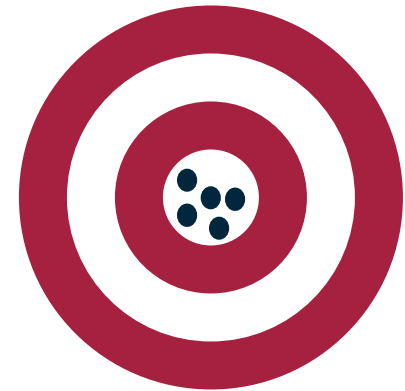


Ensemble Models



Overview

- What:
 - An ensemble model is the aggregation of two or more related but different models, averaged into a single prediction.
- Why:
 - Improve accuracy of predictions
 - Improve stability of the model



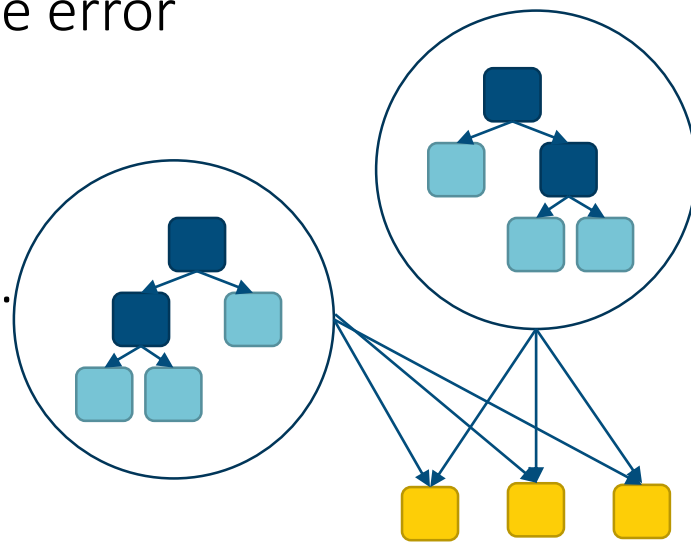
Ensemble methods

- Bagging
- Boosting
- Stacking



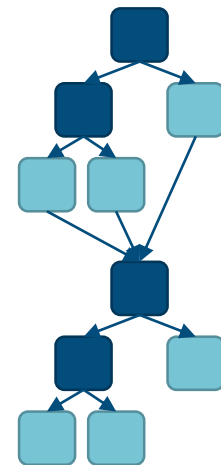
Bagging

- What is it:
 - Building multiple models from different subsamples of the training dataset, results are then combined for the final prediction.
 - Helps to reduce the variance error
- Example:
 - Random Forest
 - R package: **randomForest**, ...



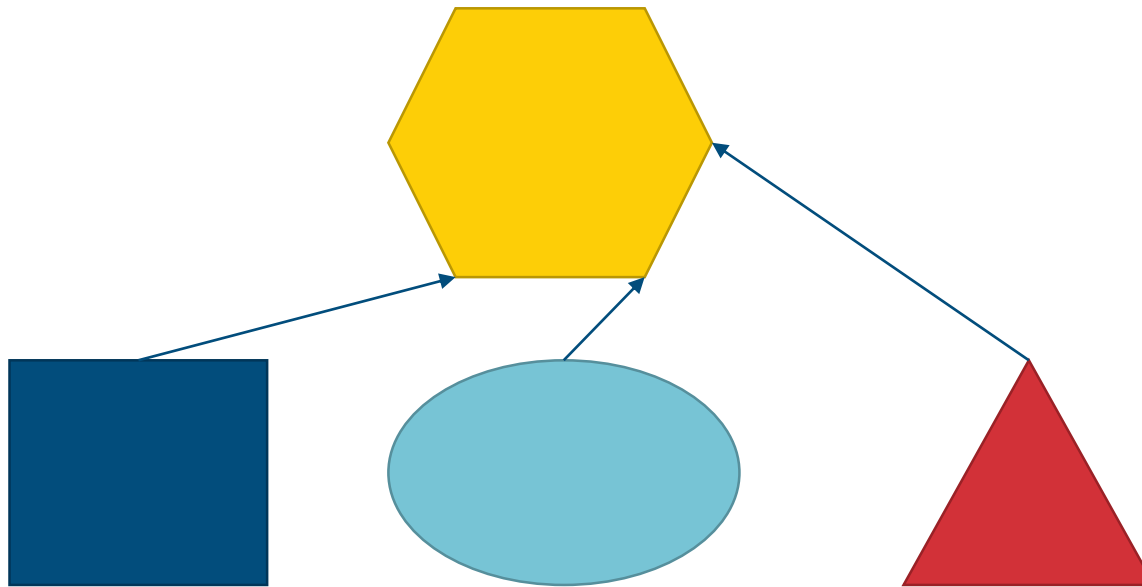
Boosting

- What is it:
 - Building multiple models, each of which is built to improve the prediction errors of a prior model
 - Has shown better predictive accuracy than bagging, but more likely to overfit
- Example:
 - Gradient Boosted Machines (GBM)
 - R packages: **gbm**, xgboost, ...



Stacking

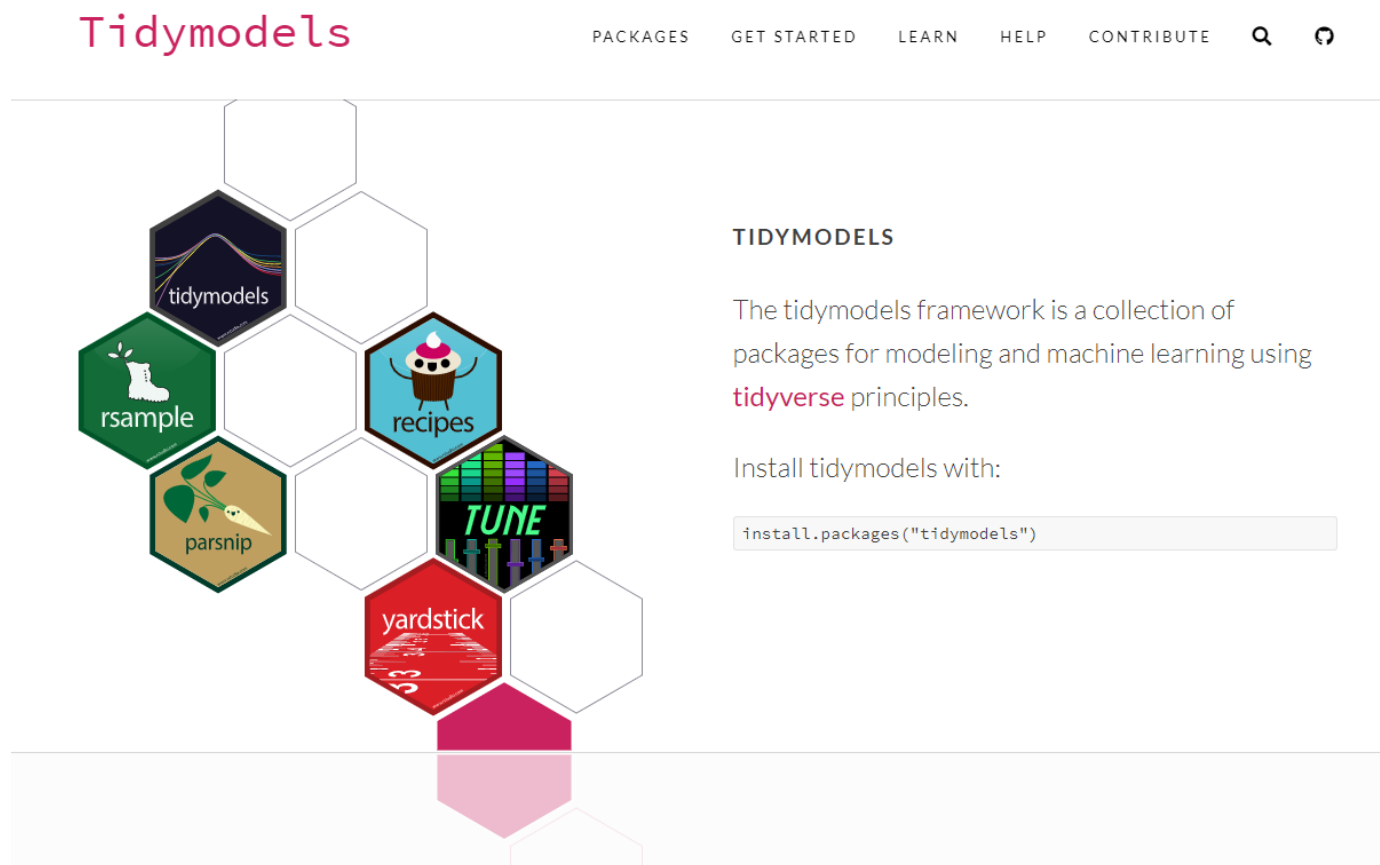
- What is it:
 - Building multiple models, typically different types of models, then having a supervisor model that determines how to best combine those results



Back to R!



Tidymodels (tidymodels.org)



The screenshot shows the homepage of the tidymodels website. At the top, the word "Tidymodels" is written in a pink, sans-serif font. To its right is a navigation bar with links: "PACKAGES", "GET STARTED", "LEARN", "HELP", "CONTRIBUTE", a search icon, and a GitHub icon. Below the navigation bar is a large graphic of a honeycomb grid. Several hexagons in the grid contain icons and names of packages: "tidymodels" (dark blue with a line graph), "rsample" (green with a boot), "parsnip" (olive with a leaf), "recipes" (light blue with a cupcake), "TUNE" (black with colorful bars), and "yardstick" (red with a ruler). To the right of the honeycomb graphic, the heading "TIDYMODELS" is followed by a paragraph: "The tidymodels framework is a collection of packages for modeling and machine learning using **tidyverse** principles." Below this, the text "Install tidymodels with:" is followed by a code block containing the command `install.packages("tidymodels")`.

Tidymodels

PACKAGES GET STARTED LEARN HELP CONTRIBUTE 🔍

TIDYMODELS

The tidymodels framework is a collection of packages for modeling and machine learning using **tidyverse** principles.

Install tidymodels with:

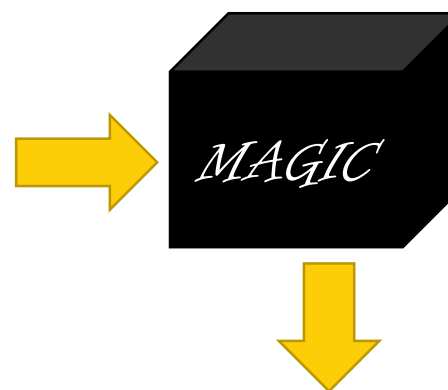
```
install.packages("tidymodels")
```


Final Thoughts



Weighing your options

- Implementation
- Explanation
- Cost



$$\text{Log Odds} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$$

Other considerations

- Actuarial judgment
- Model selection
- Data issues
- Hardware/Software



Now you're on your way!

