Confronto tra modelli di previsione

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- Tidy Modelling With R www.tmwr.org § 11.1, § 11.2, § 11.3
- van de Wiel, M.A., Berkhof, J. and van Wieringen, W.N., 2009. Testing the prediction error difference between 2 predictors. Biostatistics, 10(3), pp.550-560

Ames dataset

Il dataset `ames` contiene dati su 2930 proprietà ad Ames, Iowa, con le seguenti variabili

- caratteristiche della casa (camere da letto, garage, camino, piscina, veranda, ecc.)
- posizione (quartiere),
- informazioni sul lotto (zona, forma, dimensione, ecc.),
- valutazioni di condizione e qualità,
- prezzo di vendita.

74 variabili in tutto (40 `factor`, 22 `integer`, 12 `numeric`)

Training set: $n = 2342$ osservazioni; Test set $m = 588$ osservazioni

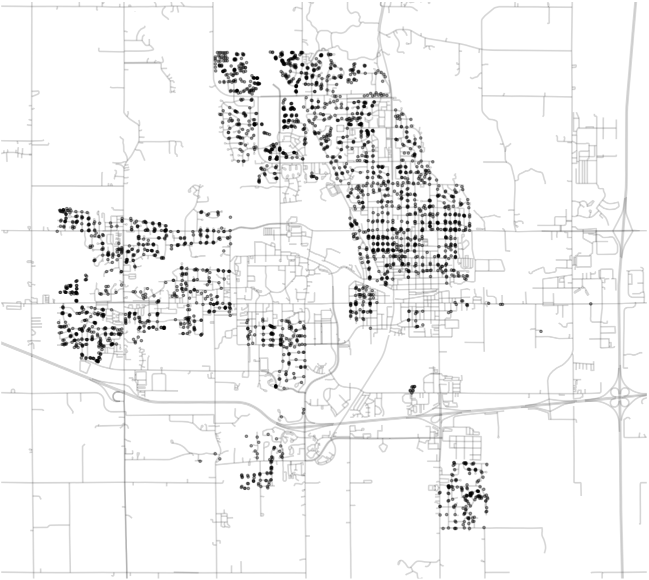


Figure 4.3 del libro *Tidy Modeling with R*: Neighborhoods in Ames, IA

Tidy Modeling with R

§ 4. The Ames housing data

§ 5. Spending our data

§ 6. Fitting models with parsnip

§ 7. A model workflow

§ 8. Feature engineering with recipes

§ 9. Judging model effectiveness

§ 10. Resampling for evaluating performance

§ 11. Comparing models with resampling

§ 12. Model tuning and the danger of overfitting

Altri riferimenti bibliografici:

De Cock (2011) https://jse.amstat.org/v19n3/decock.pdf

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Overview of *tidymodels* Basics

Package	Step	Functions
rsample	**1.** Split into testing and training sets	initial_split() training() testing()
rsample	**2.** Create recipe + assign variable roles	recipe() update_role()
recipes	**3.** Specify model, engine, and mode	parsnip function for specifying model (ex. decision_tree()) (https://www.tidymodels.org/find/parsnip/) set_engine() set_mode()
WORKFLOWS	**4.** Create workflow, add recipe, add model	workflow() add_recipe() add_model()
parsnip	**5.** Fit workflow	fit()
parsnip	**6.** Get predictions	predict()
yardstick	**7.** Use predictions to get performance metrics	rmse() (continuous outcome) accuracy() (categorical outcome) metrics() (either type of outcome)

Tidyverse Skills for Data Science: 5.13 The {tidymodels} ecosystem (v.2021-02-15)

by Carrie Wright (@mirnas22), Shannon E. Ellis (@shannon_e_ellis), Stephanie C. Hicks (@stephaniehicks), and Roger D. Peng (@rdpeng)

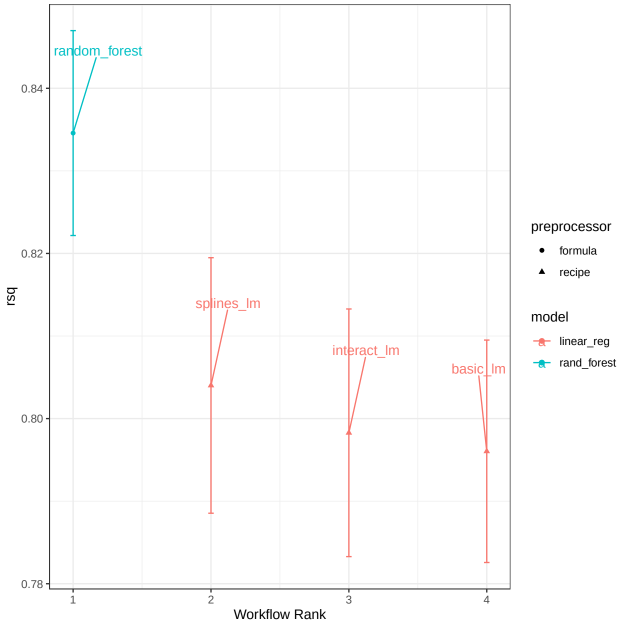
<https://jhudatascience.org/tidyversecourse/model.html#the-tidymodels-ecosystem-1>

Stima della foresta casuale

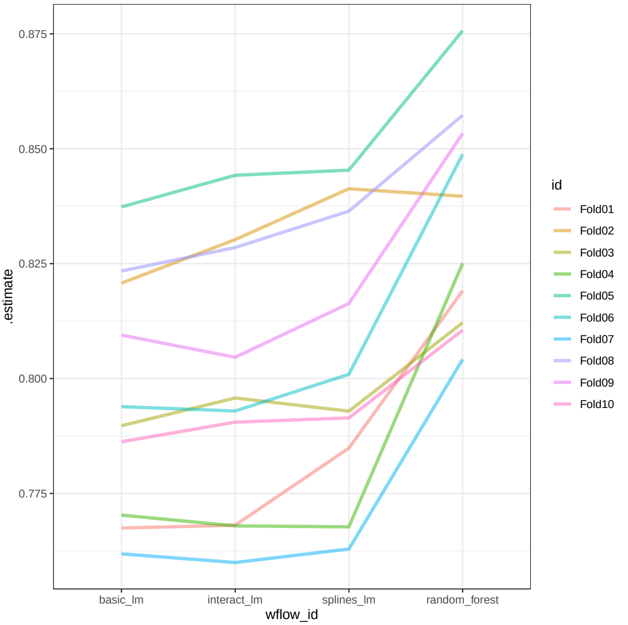
```
Sale_Price ~ Neighborhood + Gr_Liv_Area +
Year_Built + Bldg_Type + Latitude + Longitude
```

Stima dell'errore di previsione tramite convalida incrociata con 10-fold (2 fold con split 2107/235, 8 fold con split 2108/234): RMSE = 0.0720, $R^2 = 0.835$

Quattro modelli



Dieci fold



0.875
0.850
0.825
0.800
0.775
.estimate
basic_lm
interact_lm
splines_lm
random_forest
wflow_id
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

	id	rf	basic_lm	interact_lm	splines_lm	difference
1	Fold01	0.82	0.77	0.77	0.78	0.02
2	Fold02	0.84	0.82	0.83	0.84	0.02
3	Fold03	0.81	0.79	0.80	0.79	0.00
4	Fold04	0.83	0.77	0.77	0.77	-0.00
5	Fold05	0.88	0.84	0.84	0.85	0.01
6	Fold06	0.85	0.79	0.79	0.80	0.01
7	Fold07	0.80	0.76	0.76	0.76	0.00
8	Fold08	0.86	0.82	0.83	0.84	0.01
9	Fold09	0.85	0.81	0.80	0.82	0.01
10	Fold10	0.81	0.79	0.79	0.79	0.01

ANOVA

$Y = R^2$	model	X_1	X_2	X_3	id
0.8108	basic_lm	0	0	0	Fold 1
0.8134	interact_lm	1	0	0	Fold 1
0.8615	random_forest	0	1	0	Fold 1
0.8217	splines_lm	0	0	1	Fold 1
0.8045	basic_lm	0	0	0	Fold 2
0.8103	interact_lm	1	0	0	Fold 2
...					

Test della differenza di errore di previsione tra 2 modelli

Si consideri uno *split* in training $\mathcal{T}$ and validation $\mathcal{V}$.

Sui dati di training $\mathcal{T}$, stimiamo i modelli $\hat{f}_1$ e $\hat{f}_2$.

Per i dati di validation $\mathcal{V}$, otteniamo le previsioni $\hat{f}_1(x_i^*)$ e $\hat{f}_2(x_i^*)$.

Calcolare i residui $r_{i,j} = |\hat{f}_j(x_i^*) - y_i^*|$ per $j = 1, 2$ e $i \in \mathcal{V}$ e le differenze

$$d_i = r_{i,1} - r_{i,2}$$

Condizionatamente ai dati di training $\mathcal{V}$,

$$\sum_{i \in \mathcal{V}} \mathbb{1}\{d_i > 0\}$$

ha una distribuzione Binomiale di parametri $|\mathcal{V}|$ e $\pi_{\mathcal{T}} = \mathrm{pr}(|\hat{f}_1(x^*) - y^*| - |\hat{f}_2(x^*) - y^*| > 0|\mathcal{T})$. La verifica di ipotesi su $\pi_{\mathcal{T}}$, i.e. $H_0 : \pi_{\mathcal{T}} \leq 1/2$, ci permette di confrontare i due modelli