JOCLAD 2022
XXIX Jornadas de Classificação e Análise de Dados

Artificial Intelligence in Business
1st Edition – 2022

Estimates on completed buildings for the Indicators System of Urban Operations

Exploring with ML methodologies

André Sousa, António Portugal, Inês Sá, Pedro Cunha e Sara Cerdeira

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CATÓLICA LISBON
BUSINESS & ECONOMICS

18 July 2022

Summary

- Framework of the Statistical Operation
  - What is the SIOU?
  - Available data
- Estimates on Completed Buildings
  - Problem formulation
  - Calculation of estimates on completed buildings
- Alternative approaches
  - Supervised learning
- Conclusion

Framework of the Statistical Operation

What is the SIOU?

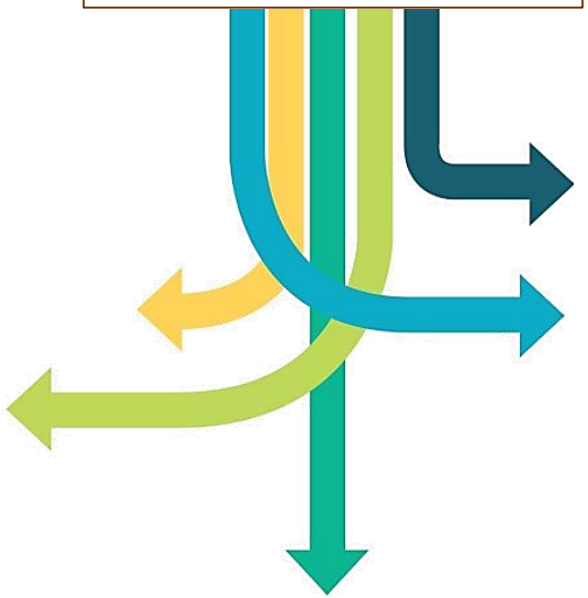
Indicators **S**ystem of **U**rban **O**perations

It is based primarily on the use of administrative information associated with the new legal framework on real estate.

Framework of the Statistical Operation

SIOU

6 surveys



Urban subdivision (S1)

Land remodeling (S2)

Change of use (S6)

Use of completed buildings (S4)

Licensing of works (S3)

Discontinued in 2017:

Conclusion of works (S5)

Framework of the Statistical Operation

What data is available from the SIOU?

Examples:

Type of work	**Destination of work**
Number of dwellings	**Territorial breakdown**
Number of floors	**Expected completion date**

Estimates on Completed Buildings

Problem:

How can we find out the conclusion date of a building?

Estimates on Completed Buildings

Mismatch between:

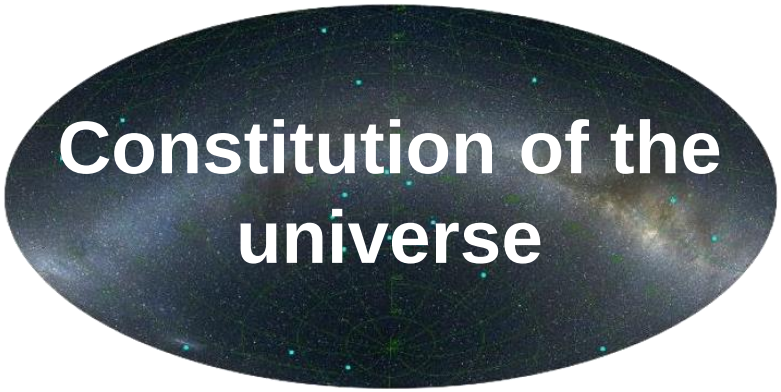
Reception of S4: Use of Completed Buildings

Data dissemination by Statistics Portugal

Solution:

Estimate the completed buildings in the period

Estimates on Completed Buildings



+ **Strata:**

Type of work

Destination of work

No. of floors

No. of dwellings

Location

Solution:

Estimate the completed buildings in the period

Estimates on Completed Buildings

Simple linear regression model (OLS):

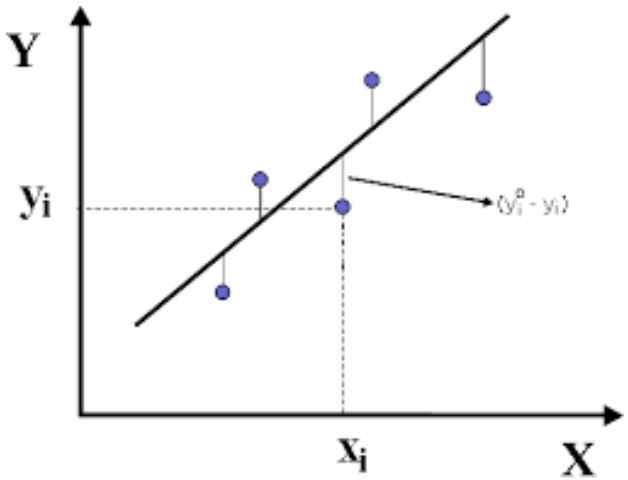
$$Y_i = \alpha + \beta X_i$$

Difference between the actual and the expected deadline (in days)

Expected deadline (in days)

Estimates on Completed Buildings

The model is estimated

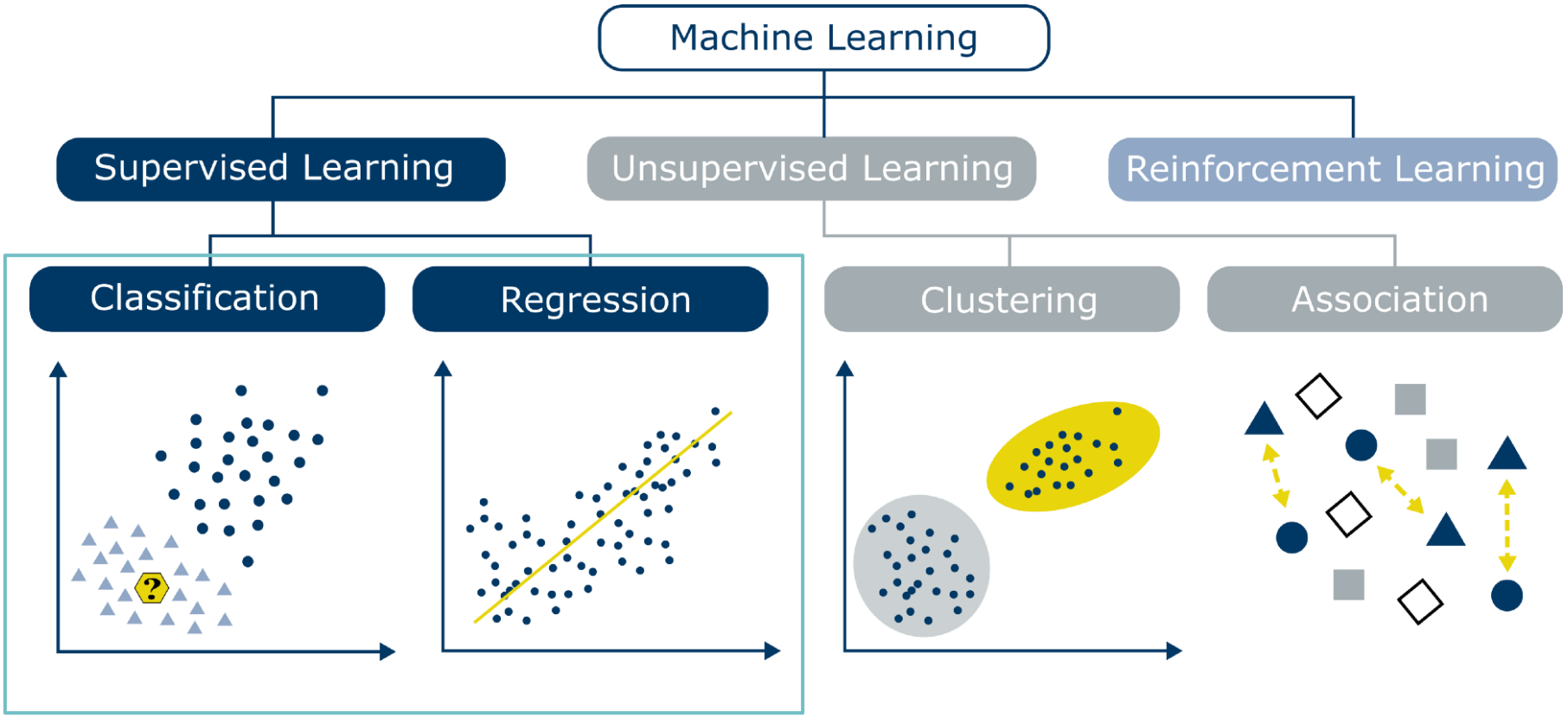


Using the least squares method

$$b = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

$$a = \bar{y} - b\bar{x}$$

Alternative approaches



Source: Dominik Polzer, 7 of the Most Used Regression Algorithms and How to Choose the Right One, Towards Data Science, Jun 21, 2021

Alternative approaches

Regression

Y = Difference between the actual and the expected deadline

Data: all records with year of completion ≥ 2011

Classification

Y = Concluded the building? (yes/no)

Data: all available records

Model	Tuned parameters	Fixed parameters	Grid
Linear/ Logistic regression	penalty mixture	-	30
Regression/ Classification tree	cost_complexity	min_n = 2 tree_depth = 30	30
Boosted trees	trees learn_rate	mtry = sqrt(k) tree_depth = 3	100 (reg) / 5 (class)

Results - Regression

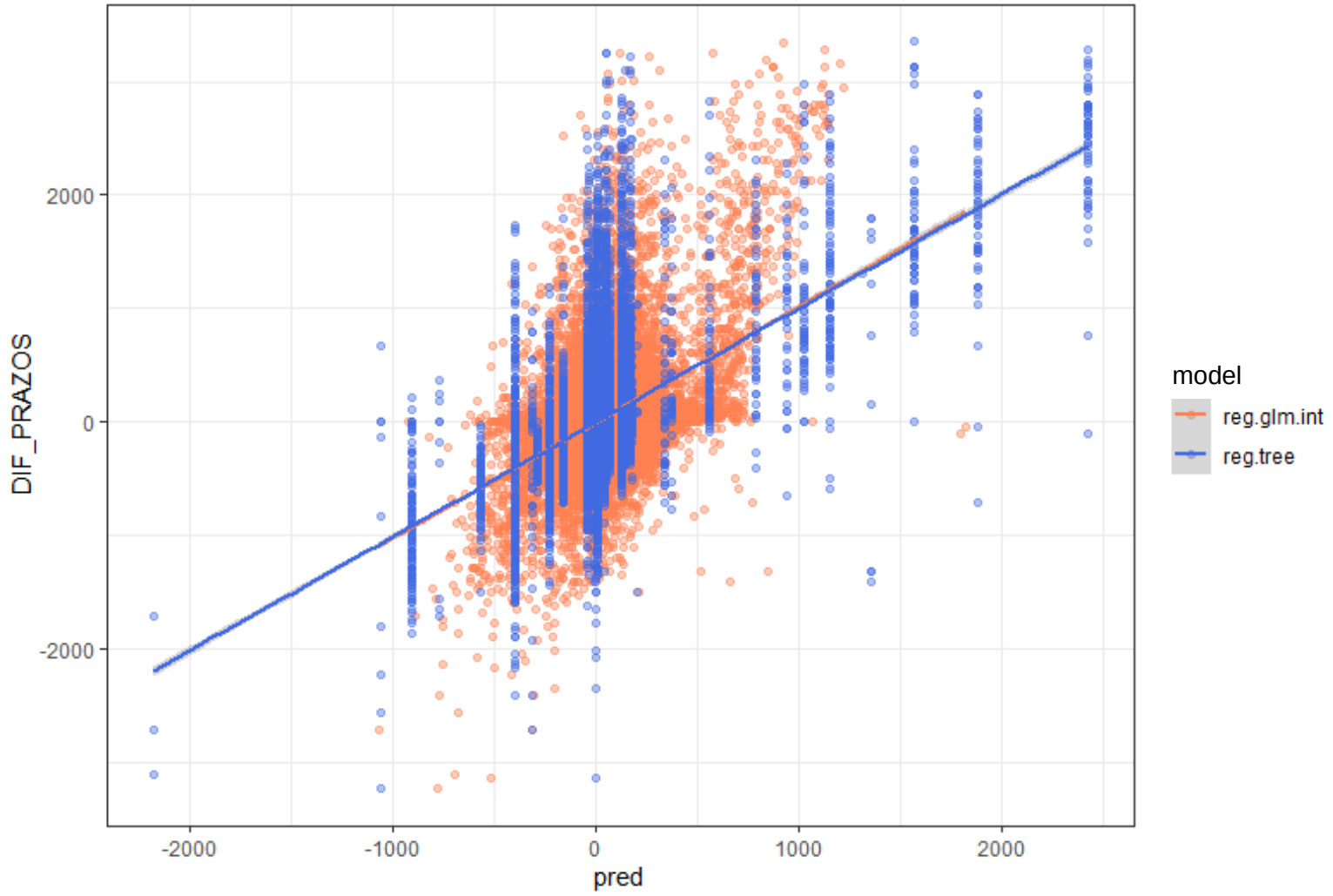
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Model	Root mean squared error (RMSE)	Mean absolute error (MAE)	Coefficient of determination (R^2)
Current model	440	258	0.0648
Linear regression w/ interaction (TO:GEO)	401	255	0.199
Regression tree	377	225	0.294
Boosted regression tree	364	219	0.341
Bagging (Reg tree + Boosted tree)	367	218	0.333

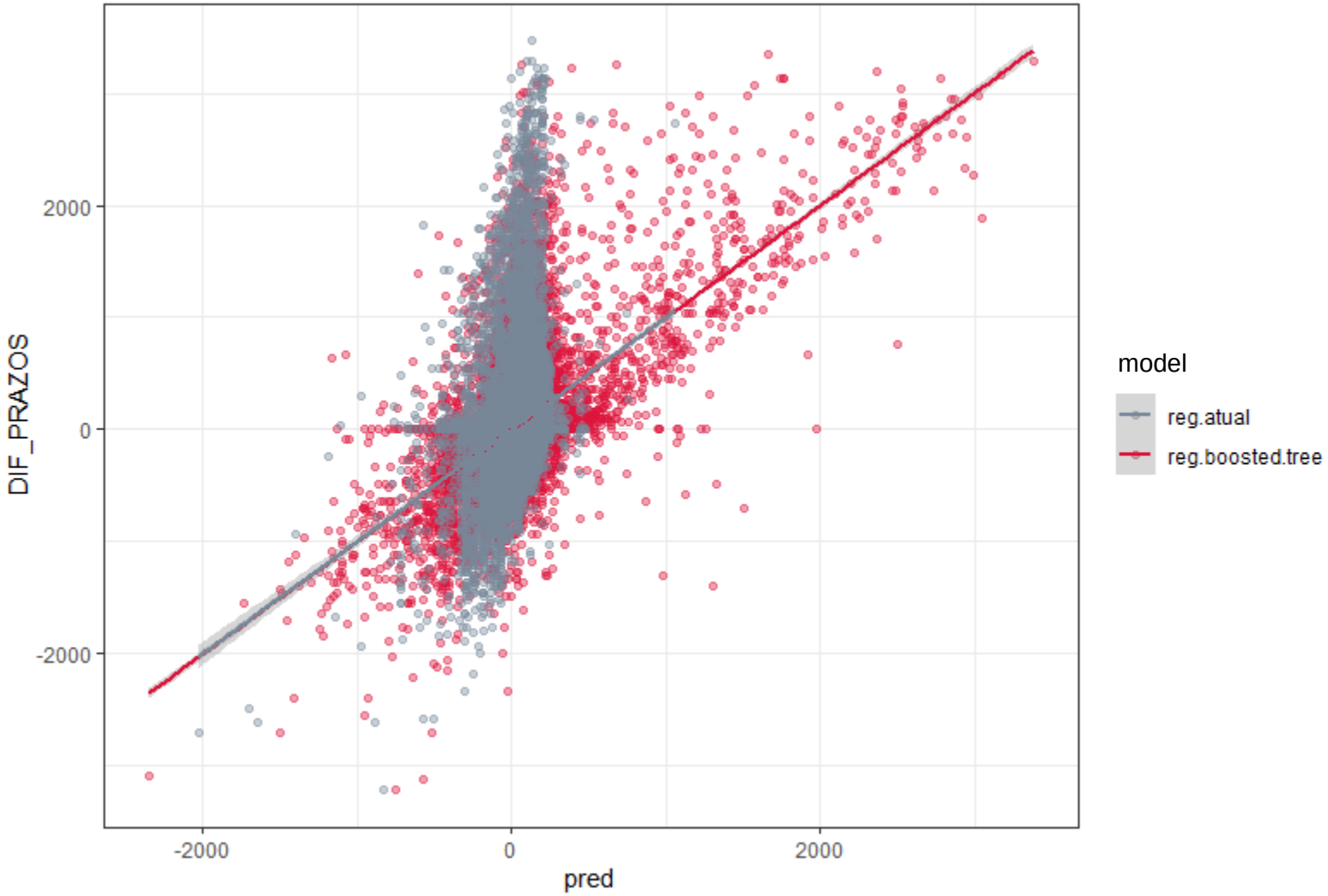
Linear regression w/ interaction (TO:GEO) vs Regression tree



Regression tree vs Boosted regression tree

Boosted regression tree vs Bagging (Reg tree + Boosted tree)

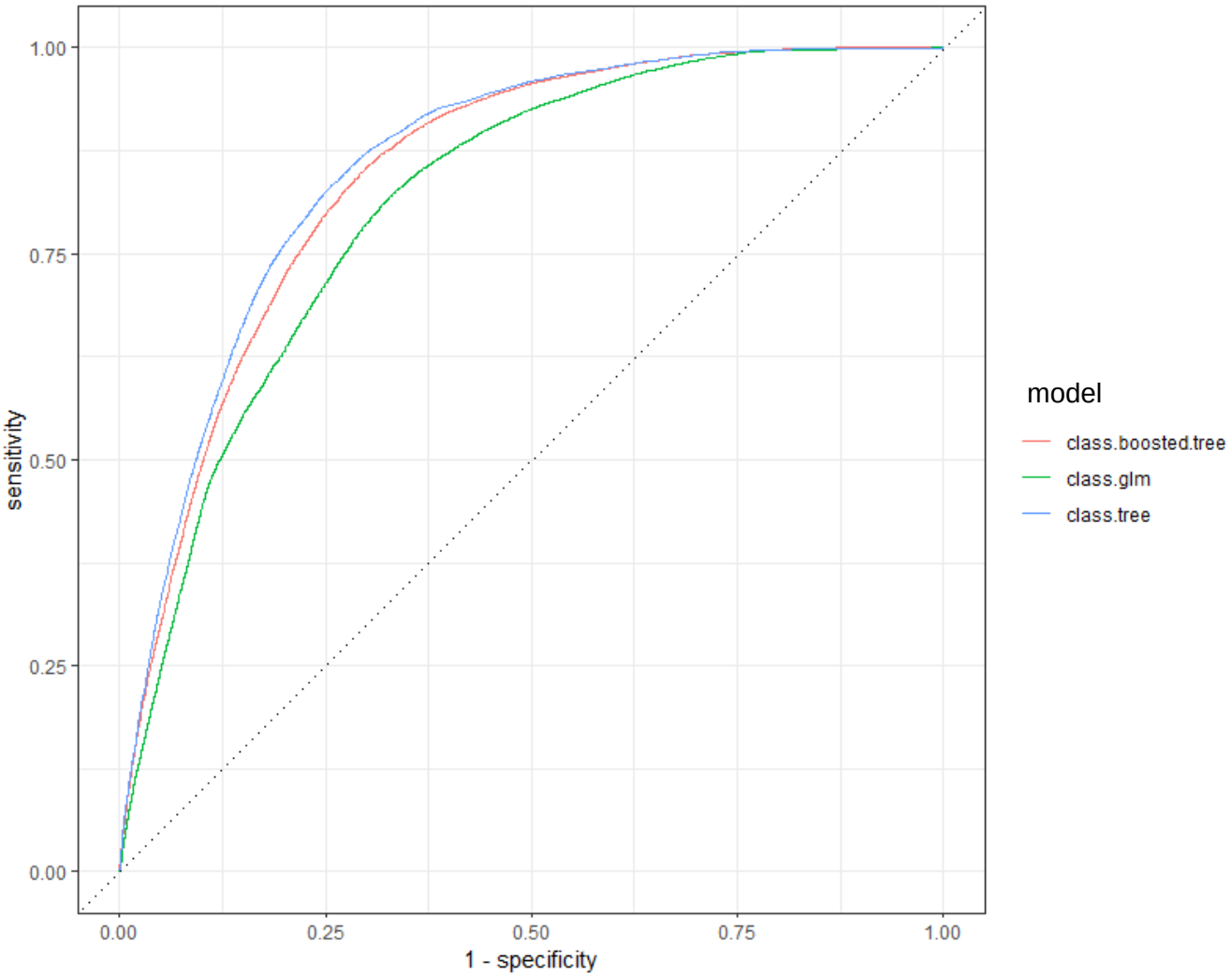
Boosted regression tree vs Current model



Results - Classification

Model	Accuracy	Sensitivity	Specificity	Precision	ROC auc
Logistic regression	0.739	0.642	0.794	0.639	0.814
Classification tree	0.787	0.726	0.821	0.697	0.858
Boosted classification tree	0.769	0.636	0.845	0.700	0.848

Results - Classification



Conclusion

- For the regression problem, the model that presented the best metrics was the **boosted regression tree** (R^2 = 0.341);
- For the classification problem, the best model was **classification tree** (ROC auc = 0.858);
- The results suggest that, in this application, the classification models are better at distinguishing the cases of completion/non-completion of the building, than the regression models at estimating the difference between the real deadline and the expected deadline for the conclusion of a building;
- Alternatives for future research: two-part model (classification + regression) or survival analysis.

JOCLAD 2022
XXIX Jornadas de Classificação e Análise de Dados

Thank you for your attention!

André Sousa
DEE/EP
andre.sousa@ine.pt

António Portugal
DMSI/II
antonio.portugal@ine.pt

Inês Sá
DMSI/ME
ines.rodrigues@ine.pt

Pedro Cunha
DMSI/II
pedro.cunha@ine.pt

Sara Cerdeira
DEE/EE
sara.cerdeira@ine.pt

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL