Estimates on completed buildings for the Indicators System of Urban Operations – Exploring with ML methodologies

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

ABSTRACT

Statistics Portugal publishes quarterly data on building permits for new constructions and on completed buildings. These data are based on the monthly permits issued by the 308 Municipalities across the country, under the scope of the Indicators System of Urban Operations (SIOU) – a system defined in 2002, based on the legislative changes resulting from the Legal Regime for Urbanization and Building. The goal behind SIOU is the gathering of data enabling us to follow the evolution of the building construction sector. It covers information on urban subdivision, land remodeling, building construction and demolition, completion of works and changes in use. The development of SIOU aimed at improving the reliability of information based on indicators and to obtain timely administrative data from the Municipal Councils, namely on the indicators relating to the completion of buildings. Given the administrative nature of the information, the data on building permits and completions are monthly updated and are subject to monthly and quarterly revisions.

Delays in receiving some information from the municipalities, namely that concerning completion of works, are at the origin of the estimation performed by Statistics Portugal concerning the completion of buildings. The difference between the real deadline and the planned deadline for the conclusion of a building is estimated based on the planned deadline (i.e., the time elapsed between the permission to build and the effective conclusion of the building, as stated in the permit), based on a linear regression model, according to geography, the building characteristics, and its final use. Such estimation method aims at decreasing deviations resulting from revisions on the quarterly estimates.

The results obtained from the current methodology were compared with those obtained by the regression problem using machine learning algorithms. Additionally, algorithms for dealing with the estimation as a classification problem were explored. Regression models and decision trees were used (namely, linear or logistic models, and regression or classification trees, depending on the approach). In addition, boosting was applied to the decision trees algorithms in an attempt to reduce prediction errors. For each model, we defined fixed values for a set of parameters and let a tuning process guide the choice of values for the remaining ones.

For the regression problem, the boosted regression tree resulted in the best predictions (RMSE = 364, MAE = 219, R^2 = 0.341); for the classification problem, the best results were obtained from the classification tree (ROC AUC = 0.858). The boosting did not improve the performance of the classification tree, but this was probably due to computing limitations that required a limited grid size for parameter searching in the tuning process.

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 planned deadline for the conclusion of a building. Alternatives for future research might include exploring a two-part

model, where the classification and the regression approaches are combined, or taking a survival analysis approach to estimate the time to conclusion subject to censoring (e.g. in cases where no information on the building completion is received from the construction promoter).

REFERENCES

James, G. et al. (2021). *An Introduction to Statistical Learning - with Applications in R*. 2nd Edition, Springer New York, NY