What Drives the Location of Data Centers? A Multi-Source Analysis of European NUTS3 Regions

Iman Zareian
Politecnico di Torino
Torino, Italy
s343587

Farshad Soofivand
Politecnico di Torino
Torino, Italy
s327536

Seyvan Azizi
Politecnico di Torino
Torino, Italy
s327960

Hossein Nabovvati
Politecnico di Torino
Torino, Italy
s343434

Ali Najafi
Politecnico di Torino
Torino, Italy
s343014

ABSTRACT

Data centers use a lot of energy and matter both for Europe's digital independence and for regional growth, but it is still not very clear what decides where they end up. We study this question on a dataset of 1,345 European NUTS3 regions that puts together socio-economic, infrastructure, and environmental variables. We treat the problem as binary classification — does a region have at least one data center or not? — and train two models side by side: a logistic regression with `statsmodels`, which is easy to interpret, and a tuned Random Forest with `scikit-learn`, which focuses on prediction. Both reach a test ROC-AUC of about 0.77, well above the 0.559 majority baseline, and both agree on the main point: a region's economic size, measured by GDP, matters most, raising the odds of having a data center about four times per standard deviation. Connectivity has a weaker effect in the expected direction, and the direct climate variables do not seem to matter at this level of detail. One result was surprising at first: population had a negative coefficient in the logistic model, but a SHAP analysis showed this was an artefact of its correlation with GDP, not a real negative effect.

KEYWORDS

Data centers, NUTS3 regions, Logistic regression, Random Forest, SHAP, Multicollinearity, Digital infrastructure

1 INTRODUCTION

Data centers are the physical backbone of the digital economy. They run the cloud services, the storage, and the growing amount of AI work that everything else depends on, and where they are built is not random. Companies look at how well a place is connected, how close it is to customers and capital, how cheap and reliable the electricity is, and — more and more — the local climate, because cooling is a big part of the operating cost. For the European Union, which wants to be more digitally independent and reach net-zero at the same time, knowing where data centers cluster and why is a practical question, not only an academic one.

We try to answer one question: *which socio-economic, infrastructure, and environmental variables best explain where data centers are present across European NUTS3 regions?* From the literature we formed three hypotheses:

- **H1 (Connectivity).** Regions with more Internet Exchange Points (IXPs) and better network performance are more likely to have data centers.
- **H2 (Economic mass).** Regions with higher GDP and population are more likely to have data centers.
- **H3 (Energy and climate).** Cheaper electricity and a milder climate make a region more likely to have a data center.

The rest of the paper follows the data–model–interpretation approach used in class: clean the data without leaking test information, train an easy-to-interpret model to find the drivers, train a stronger model to get better predictions, and use SHAP to interpret it.

2 RELATED WORK

Data centers in Europe are very unevenly spread. The Frankfurt–London–Amsterdam–Paris cluster, later joined by Dublin, grew because of strong connectivity, closeness to big markets, and good energy and regulatory conditions. Studies of cloud infrastructure often put energy cost and reliability among the most important reasons for choosing a location, and regional studies at the NUTS level link high-value economic activity to GDP and the size of the local workforce [2]. For the methods, we use standard tools: logistic regression for interpretation, the Random Forest [1] for prediction, and SHAP [3] to explain it, all using `scikit-learn` [4] and `statsmodels` [5]. What we add is to put these variables together at the finer NUTS3 level and to measure how much each one really matters, using an interpretable model and a flexible one side by side.

3 DATA

Our data cover 1,345 NUTS3 regions, each described by 45 columns. These belong to three groups: socio-economic (GDP, population), infrastructure (IXP count and distance, broadband download and upload speed, latency, power-plant capacity), and environmental (solar radiation, wind speed, drought indices, climate-related losses, pollutant releases). The original target is the number of data centers in a region, and this count is very uneven — most regions have zero, while a few have many (max 106).

To make the target stable, we converted it to a binary label `has_datacenter`, equal to 1 when a region has at least one center. About 44% of regions are positive, so a model that always predicts "no data center" is already right 55.9% of the time — the baseline

Table 1: Test-set performance ($n = 269$). Both models clearly beat the majority baseline, and the two are statistically indistinguishable from each other.

Model	Accuracy	F1	ROC-AUC
Baseline (majority class)	0.559	—	0.500
Logistic Regression (statsmodels)	**0.706**	**0.672**	0.766
Random Forest (scikit-learn)	0.699	0.658	**0.774**

Table 2: Significant logistic-regression drivers ($p < 0.05$). Odds ratios are per one standard deviation of the standardized predictor.

Feature	Coef.	Odds ratio	p
`gdp_eur_million`	+1.444	4.24	< 0.001
`Air Releases`	+0.319	1.38	0.021
`population`	−0.409	0.66	0.048

our models have to beat. The data also have missing values (up to 15.5% for the mortality variables) and several variables that overlap or are nearly the same, so cleaning and feature selection were an important part of the work, not an afterthought.

4 METHODOLOGY

Problem framing. We treat this as classification. The count is too skewed for regression to be stable, and the binary version answers the question we really care about — what makes a data center present — more directly.

Preprocessing and feature selection. Before computing anything we split the data into a stratified 80% training set ($n = 1076$) and a 20% test set ($n = 269$). This is important: everything we estimate afterwards — the medians used for imputation, the features we drop, the mean and variance used for scaling — comes only from the training set and is then applied to the test set, so no test information leaks back. We filled missing values with training medians. Feature selection had two steps. First we dropped one variable from each pair with correlation above 0.85, which removed 10 redundant features. But pairwise correlation does not catch collinearity that appears across several variables together: a few features still had a Variance Inflation Factor (VIF) above 10, mainly different wind-speed statistics. So we used an iterative VIF step, dropping the feature with the highest VIF and recomputing until all were below 10, which removed 2 more (`wind_min_speed_100m` and `ixp_total_participants`). We then log-transformed the skewed non-negative variables and standardized the rest. This left 27 features, all with VIF below 7.2.

Models. We trained two models on the same features, as the assignment asks. The interpretable one is a logistic regression fitted with `statsmodels`, where we read coefficients, odds ratios, and p-values to see what drives the outcome. The prediction one is a Random Forest from `scikit-learn`, with its main hyperparameters (`max_depth`, `min_samples_leaf`, `max_features`, `class_weight`) tuned by 5-fold stratified cross-validation on ROC-AUC. We fixed `random_state = 42` everywhere so the results can be reproduced.

Evaluation and interpretation. We report accuracy, F1, and ROC-AUC on the test set, always compared with the majority baseline. To understand the Random Forest beyond accuracy, we use permutation importance (how much the test ROC-AUC drops when a feature is shuffled) together with SHAP, which also shows the direction in which each feature pushes a prediction.

5 RESULTS

Table 1 shows the test-set results. Both models clearly beat the baseline, and they perform almost the same as each other: the Random Forest is a little better on ROC-AUC, while the simpler logistic model is a little better on accuracy and F1.

The logistic model fits well, with a McFadden pseudo-R^2 of 0.297 and a likelihood-ratio test that is clearly significant ($p < 10^{-75}$). Three predictors pass the $p < 0.05$ bar (Table 2). Because the features are standardized, each odds ratio is the effect of a one-standard-deviation increase. GDP is far ahead of the rest, multiplying the odds of having a data center by 4.24.

For the Random Forest the cross-validation ROC-AUC was 0.828, a bit higher than the 0.774 on the test set, which is normal for a small dataset. What matters is that the model works evenly on both classes (precision and recall around 0.73 for the majority class and 0.66 for the minority), so it is not just predicting the bigger class every time.

The two ROC curves (Figure 1) are almost on top of each other, which gives the same message as Table 1. Permutation importance gives the same picture for GDP: shuffling it lowers the test ROC-AUC by 0.058, about ten times more than any other feature (the rest are near 0.005). The SHAP summary in Figure 2 adds the direction. GDP always increases the predicted probability; variables that act as proxies for economic and industrial activity, such as broadband test volume and pollutant releases, also push it up; connectivity variables like latency and download speed have weaker but reasonable effects; and the climate variables stay close to zero.

6 DISCUSSION

H2 (economic mass): strongly supported. GDP is the strongest driver in every measure we looked at — the logistic coefficient, permutation importance, and SHAP all agree. A region's economic size explains its data centers better than any infrastructure or environmental variable.

Reconciling the population paradox. The logistic model gave population a negative and significant coefficient, which at first looked like it went against H2. SHAP explains it: in the Random Forest, population's effect is clearly positive (Figure 2). The negative linear coefficient is a side effect of collinearity — once GDP, which moves together with population, is in a linear model, the remaining population term takes a misleading negative sign. The tree model does not force a straight-line relationship, so it recovers the positive effect we would expect. This is a good example of why it helps to use an interpretable model together with a flexible one: each catches what the other misses.

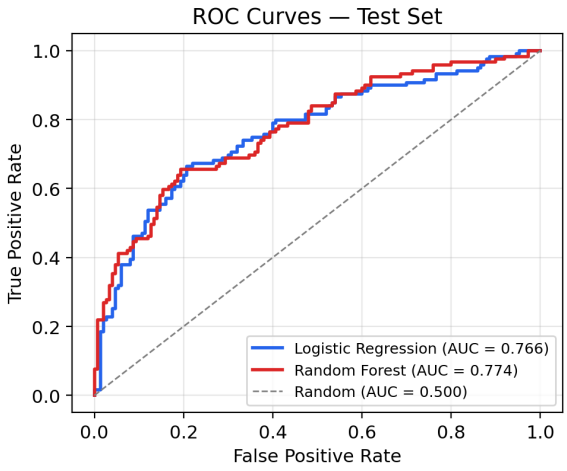

Figure 1: ROC curves on the held-out test set. The two models are indistinguishable (ROC-AUC ≈ 0.77), both well above the random baseline.

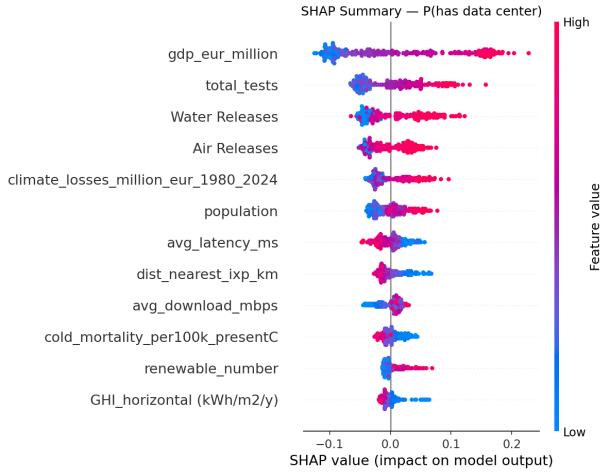

Figure 2: SHAP summary for the Random Forest. GDP dominates and always pushes the prediction up; industrial-activity proxies contribute positively; and population's effect is positive here, contrary to its negative logistic coefficient.

H1 (connectivity): a weak secondary effect. IXP count and lower latency point in the right direction and are among the more useful features, but they just miss significance in the logistic model and do not change predictions much. Connectivity matters, but it is a second-order factor once economic size is taken into account.

H3 (energy and climate): not supported. Electricity price, renewable capacity, solar radiation, and the drought indices are neither significant nor important in either model. Whatever effect climate has on location, it is not visible at this level of detail.

Model comparison and limitations. The fact that the Random Forest only matches the logistic model suggests the relationship is mostly linear and mostly about GDP, so the extra flexibility does not add much. A few limitations should be stated clearly. The data are from a single point in time, so we describe association, not cause. Nearby regions are probably related in ways our models ignore. And turning the count into a yes/no label loses information about how many data centers a region has.

7 CONCLUSION

We modelled the presence of data centers across 1,345 European NUTS3 regions using a logistic regression and a tuned Random Forest. Both reach a test ROC-AUC near 0.77, and both agree that a region's economic size, measured by GDP, is the main thing that decides where data centers are, with connectivity a distant second and climate showing no clear effect. SHAP also helped us correct a misleading result in the linear model, showing that population is really positively linked to data-center presence once collinearity is taken into account. Good next steps would be to use models that take spatial structure into account, to use data over time instead of one snapshot, and to predict the data-center count directly so that the amount, not only the presence, is part of the analysis.

REFERENCES

[1] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
[2] Eurostat. Regional Statistics by NUTS Classification. https://ec.europa.eu/eurostat/web/regions, accessed 2026.
[3] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, volume 30, pages 4765–4774, 2017.
[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[5] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In *SciPy*, pages 92–96, 2010.