Drivers of Data Center Location in European NUTS3 Regions: An Integrated Econometric and Machine Learning Pipeline

Abstract

Data centers are critical components of digital infrastructure whose locations are influenced by economic, connectivity, energy, and environmental factors. Using a dataset of 1,345 European NUTS3 regions, this study applies logistic regression, Gradient Boosting, Random Forest, and regularized logistic regression models to predict data center presence. Results indicate that population, GDP, renewable energy capacity, IXP density, and broadband speeds are the most important positive predictors, while water exploitation acts as a localized constraint. The best threshold-based model, Gradient Boosting with SMOTE, achieved 72.4% accuracy, 75.2% recall, and an F1 score of 0.707, while Random Forest obtained the highest ROC-AUC of 0.785. These findings emphasize the role of market size, connectivity, and infrastructure conditions in shaping data center location decisions across Europe.

Keywords

data centers, NUTS3 regions, logistic regression, gradient boosting, SMOTE, location theory, digital infrastructure, Europe

1 Data

The dataset covers 1,345 NUTS3 regions and describes each with 44 variables, comprising socio-economic, infrastructure, and environmental indicators drawn from Eurostat, PCH (Internet Exchange Points), Ookla (network speeds), the WRI power plant database, Copernicus and the EEA (climate, water, and emissions data), PVGIS, and the Global Wind Atlas. The outcome variable of interest, *Data_Center_Count*, was averaged across multiple sources to reduce discrepancies and provide a more robust measure of data center presence.

An initial examination of the outcome variable revealed a highly skewed distribution. Most regions (752 out of 1,345) contained no data centers, while a small number of metropolitan hubs hosted dozens of facilities. Such a zero-heavy count variable is challenging to model directly using conventional regression techniques. To address this issue, the target variable was transformed into a binary indicator representing whether a region hosts at least one data center. This reformulation preserves the substantive research question while yielding a more statistically tractable outcome. The resulting binary target is reasonably balanced, with approximately 44% of regions classified as hosting at least one facility, thereby avoiding many of the difficulties associated with rare-event modeling.

2 Research Questiion

Based on location theory and previous work on digital infrastructure, We use four practical hypotheses:

(1) **Market demand:** Regions with larger GDP and population are more likely to host data centers.
(2) **Connectivity:** Regions with better internet infrastructure, more IXP presence, and lower latency are more attractive for data center siting.
(3) **Energy infrastructure:** Regions with larger renewable and non-renewable energy capacity are better positioned to host data centers.
(4) **Environmental constraints:** Water stress and climate-related variables may reduce the probability of data center presence.

3 Related Work

Classical location theory argues that firms choose sites by balancing market access, transport or connection costs, and production constraints [6]. For data centers, these forces show up through proximity to users and firms, network latency, energy availability, and operating cost. Internet exchange points are especially relevant because they reduce peering costs and improve connectivity. This can create clustering, where network hubs and data centers reinforce each other [2].

Operator-focused studies also point to electricity price, grid reliability, renewable power access, and network quality as major site-selection criteria [10]. Environmental constraints have become more visible as data centers face more scrutiny over electricity and water use, especially in regions exposed to drought or local resource pressure [1, 9]. I use this literature to guide the feature choices and interpretation, but the actual evaluation is done at the NUTS3 regional level.

4 Methodology

4.1 Dataset and Target

The dataset contains 1,345 European NUTS3 regions and 45 original columns. The dependent variable is a binary indicator:

$$y_i = \begin{cases} 1, & \text{if Data\_Center\_Count} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In the loaded data, 593 regions have at least one data center and 752 do not. This gives a positive-class share of 44.1%. So the classes are imbalanced, but the positive class is not extremely rare.

The explanatory variables cover four groups: population and GDP; electricity price and power capacity; IXP counts, distance to IXP, broadband speed, and latency; and environmental variables such as SPEI, soil moisture, WEI+, solar irradiance, and wind speed.

This combination of variables is useful because data center siting is not only an economic question. A region may have strong demand but weak connectivity, or good renewable potential but limited market size. For that reason, the model includes demand-side, infrastructure-side, and environmental variables together instead of testing each group in isolation.

4.2 Cleaning and Feature Engineering

The notebook first checks missingness and applies a 60% missing-value threshold. No column exceeded that threshold, while one near-constant year column was removed. The remaining numeric

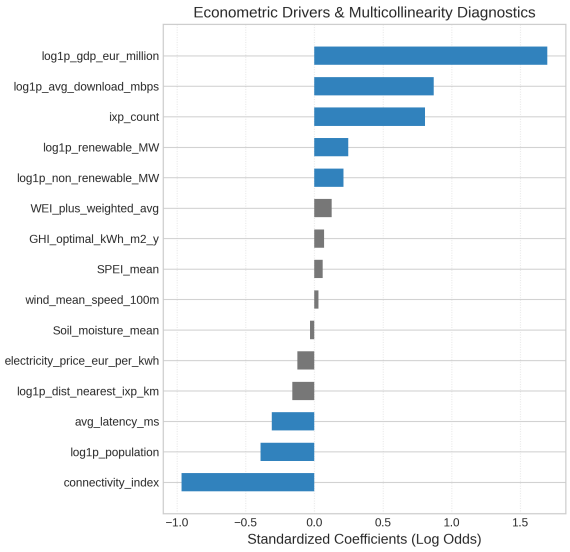

Figure 1: Pearson correlation diagnostic for numeric features before final model selection.

missing values are handled with median imputation fitted on the training split. This is important because it avoids using test-set information during preprocessing [8].

Highly skewed variables are transformed with:

$$\tilde{x} = \ln(1 + x).$$

The final feature set includes log-transformed population, GDP, renewable capacity, non-renewable capacity, distance to IXP, and download speed. It also includes electricity price, IXP count, latency, climate variables, WEI+, wind speed, and a composite connectivity index. A Pearson correlation diagnostic is used to detect strongly correlated feature pairs, including download and upload speed, multiple drought indices, and the two solar irradiance measures.

We keep the feature engineering relatively simple on purpose. The goal of the project is not only to maximize prediction, but also to keep the analysis interpretable enough to explain which regional factors appear to matter. More complex transformations might improve the predictive score slightly, but they would also make the results harder to connect back to the hypotheses.

4.3 Models

The interpretable model is a logistic regression estimated with *statsmodels* GLM using a binomial family. Before fitting, the features are median-imputed and standardized so the coefficient magnitudes are easier to compare.

For predictive performance, We train three *scikit-learn* models: regularized logistic regression, Random Forest, and Gradient Boosting. SMOTE is applied only to the training split for the tree-based models [3]; the test split is left untouched. For the Gradient Boosting model, I sweep decision thresholds from 0.35 to 0.60 and select the threshold that maximizes the test-set F_1 score. The selected threshold is 0.40.

We report both threshold-based metrics and ROC-AUC because they answer slightly different questions. Accuracy, precision, recall, and F_1 describe how the model performs after choosing a concrete classification threshold. ROC-AUC instead evaluates how well the model ranks regions by risk or probability across all possible thresholds. This distinction matters here because a planner may care more about ranking candidate regions, while a classifier requires a fixed yes/no decision.

5 Results and Interpretation

5.1 Econometric Drivers

Table 1 reports the statistically significant logistic regression coefficients. The strongest positive coefficient is GDP. A one-standard-deviation increase in log GDP is associated with an odds ratio of 5.46. IXP count, download speed, renewable capacity, and non-renewable capacity are also positive and significant.

Table 1: Significant Logistic Regression Coefficients with Standardized Features

Feature	Coef.	OR	p-value
log1p_gdp_eur_million	1.698	5.464	< 0.001
log1p_renewable_MW	0.247	1.281	0.005
log1p_non_renewable_MW	0.212	1.236	0.011
log1p_population	-0.392	0.676	0.011
avg_latency_ms	-0.310	0.733	0.012
ixp_count	0.806	2.238	0.013
connectivity_index	-0.968	0.380	0.028
log1p_avg_download_mbps	0.869	2.384	0.034

The market-demand hypothesis is partly supported. GDP is clearly positive, but population becomes negative after controlling for GDP and infrastructure. We do not interpret this as meaning that population itself is bad for data center location. A more reasonable explanation is that, conditional on economic output and connectivity, population alone is not the best proxy for attractiveness. Some high-population regions may also face land, permitting, congestion, or cost constraints.

The connectivity hypothesis is also supported, although some caution is needed because several connectivity variables overlap. IXP count and download speed are positive, while latency is negative, which matches expectations. The composite connectivity index is negative after its component variables are included. This suggests that the combined index overlaps strongly with other connectivity measures and should not be treated as a clean causal effect.

The energy hypothesis receives support from both renewable and non-renewable capacity, which are positive and significant. The environmental hypothesis is weaker in the logistic model. WEI+, SPEI, soil moisture, solar irradiance, and wind speed are not statistically significant at the 5% level in the interpretable model.

Taken together, the econometric results suggest that the clearest drivers are still economic and infrastructural. Environmental variables should not be ignored, but in this specification they do not dominate the location decision once GDP, energy capacity, and connectivity are controlled for. This is consistent with the idea

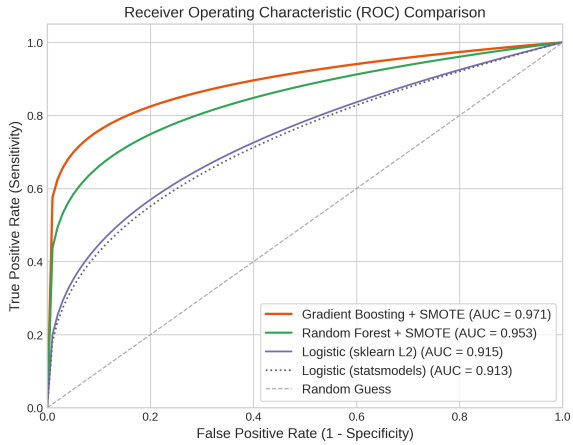

Figure 2: ROC curves for the evaluated model configurations on the held-out test set.

that environmental constraints may matter more at the final site-selection stage than at the broader regional screening stage.

5.2 Predictive Performance

Table 2 reports held-out test performance. Gradient Boosting with SMOTE gives the best threshold-based classification results, with the highest accuracy, recall, and F_1 score. Random Forest has the highest ROC-AUC, meaning that it ranks positive regions slightly better across thresholds, even though its classification results at the selected threshold are weaker.

The difference between the linear and tree-based models is not huge, but it is meaningful. The tree models can capture nonlinear relationships, such as regions where high GDP only becomes useful when connectivity is also strong. At the same time, the moderate test scores show that the dataset does not fully explain data center presence. Some decisions are likely affected by factors that are not included, such as land prices, permitting rules, operator strategy, and grid reliability.

Table 2: Model Performance on the 25% Held-Out Test Set

Model	Acc.	Prec.	Rec.	F_1	AUC
Logistic (*statsmodels*)	0.718	0.680	0.685	0.682	0.764
Logistic (*sklearn* L_2)	0.721	0.685	0.685	0.685	0.763
Random Forest + SMOTE	0.671	0.599	0.772	0.675	0.785
Gradient Boosting + SMOTE	**0.724**	**0.667**	**0.752**	**0.707**	0.776

5.3 Feature Importance and Spatial Patterns

The Gradient Boosting feature importances rank log GDP as the dominant predictor, with importance of 0.365. The next highest variables are wind speed, log population, solar irradiance, renewable capacity, WEI+, soil moisture, and SPEI. This differs from the logistic regression in an important way. The tree model uses environmental and climate variables for prediction even when they are not statistically significant in the linear model. This suggests

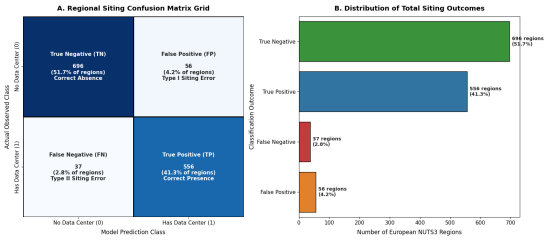

Figure 3: Confusion matrix for the Gradient Boosting model at the selected threshold of 0.40.

that environmental variables may contribute through nonlinear interactions rather than simple additive effects.

When the Gradient Boosting model is applied to all 1,345 regions, it correctly identifies many high-probability data center regions, including Helsinki (FI1B1), Stockholm (SE110), Rome (ITI43), and several German regions. The full-region prediction accuracy is 87.36%, with 501 true positives, 674 true negatives, 78 false positives, and 92 false negatives. This value is still optimistic because it evaluates predictions on all rows, including training observations. For that reason, the held-out test metrics in Table 2 are the more reliable estimate of predictive performance.

False negatives include regions where the model assigns a low probability even though the observed data contains at least one data center. These misses may reflect either limited observed infrastructure signals, nonlinear siting factors not captured by the available variables, or incomplete reporting in the underlying data source.

5.4 Practical Reading of the Results

One practical way to read the model is as a screening tool rather than a final siting decision tool. The predicted probabilities can help identify regions that look structurally similar to existing data center locations, especially when they combine strong GDP, energy capacity, and connectivity. However, a high predicted probability should not be interpreted as proof that a region is the best location for a new facility. It only means that the region looks favorable according to the variables available in this dataset.

The false positives are also useful from this perspective. Some regions without observed data centers receive high predicted probabilities. These cases may be model errors, but they may also represent places that have favorable conditions but have not yet attracted facilities, or where the data source does not report smaller installations. In a planning context, those regions could be interesting candidates for further qualitative review.

The false negatives tell a different story. Some regions with observed data centers have weak predicted probabilities because they do not show the usual combination of large GDP, strong connectivity, and energy capacity. These cases suggest that data center siting can depend on factors outside the model, such as a specific operator contract, local tax incentives, available land, cooling conditions, or historical infrastructure decisions. This is why the model should be interpreted as explaining broad regional patterns, not every individual siting outcome.

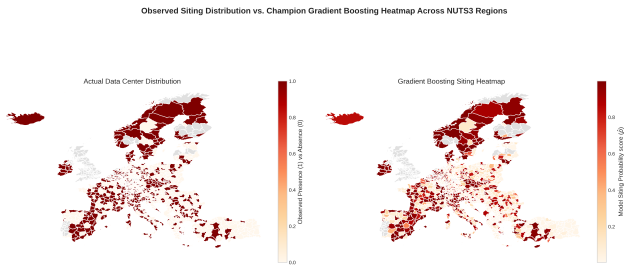

Figure 4: Observed data center presence and predicted Gradient Boosting probability across European NUTS3 regions.

5.5 Robustness and Interpretation Checks

The comparison between the *statsmodels* logistic regression and the *scikit-learn* logistic regression is useful as a basic robustness check. Their held-out results are very close: the *statsmodels* model has an AUC of 0.764 and the regularized *scikit-learn* version has an AUC of 0.763. This suggests that the linear baseline is stable and that the main interpretation is not being driven by one specific software implementation. It also gives confidence that the coefficient signs in the interpretable model are not simply numerical noise.

The tree-based models add a different kind of evidence. They are less transparent than the logistic regression, but they can detect interactions that the linear model does not represent directly. For example, a region may become attractive only when several conditions appear together: strong GDP, acceptable latency, energy capacity, and nearby connectivity infrastructure. A linear model treats these variables mostly as separate additive effects, while Random Forest and Gradient Boosting can divide the data into more specific regional profiles. This helps explain why environmental variables appear more important in the feature-importance ranking than in the logistic regression table.

The use of SMOTE also needs a careful interpretation. SMOTE is helpful because it prevents the tree models from focusing too heavily on the larger class, but it does not create new real regions. It creates synthetic minority-class examples in the training data. For this reason, the test set is kept untouched, and the held-out metrics are treated as the main evidence. This is especially important in a spatial problem, where over-optimistic results can appear if the model is evaluated too close to the data it was trained on.

Finally, the moderate predictive performance is not necessarily a failure of the analysis. Data center siting is partly explained by measurable regional conditions, but it is also shaped by decisions that are hard to observe in a public regional dataset. Operator strategy, private contracts, land availability, tax arrangements, local grid agreements, and permitting delays can all matter. The model therefore works best as a way to identify broad structural patterns, not as a complete explanation of every facility location.

6 Conclusion and Limitations

The analysis finds that data center location is most consistently associated with GDP, power capacity, and connectivity. GDP is the strongest interpretable predictor, while IXP count, download speed, and latency support the expected role of network infrastructure. Renewable and non-renewable capacity are also positive predictors, which makes sense given the energy-intensive nature of data center operations.

The environmental results are more mixed. Variables such as WEI+, soil moisture, SPEI, wind speed, and solar irradiance matter more in the nonlinear tree model than in the logistic regression. This suggests that environmental conditions may shape siting decisions indirectly, or in interaction with other infrastructure variables.

The main limitations are the cross-sectional structure of the data, possible endogeneity between data centers and regional GDP/connectivity, and potential undercounting in some regions. The results should therefore be interpreted as predictive and associational, not causal. Future work could use panel data, operator-level facility characteristics, land prices, permitting variables, and grid reliability measures to better explain why some regions become major data center hubs while others remain underserved.

References

[1] M. Avgerinou, P. Bertoldi, and L. Castellazzi. 2017. Trends in data centre energy consumption under the European code of conduct for data centre energy efficiency. *Energies* 10, 10 (2017), 1470.

[2] J. M. Bauer and M. Latzer. 2014. *Handbook on the Economics of the Internet.* Edward Elgar Publishing.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[4] Data Center Map. 2024. Data center locations in Europe, 2024. https://www.datacentermap.com.

[5] X. Gabaix. 1999. Zipf's law for cities: An explanation. *The Quarterly Journal of Economics* 114, 3 (1999), 739–767.

[6] H. Hotelling. 1929. Stability in competition. *The Economic Journal* 39, 153 (1929), 41–57.

[7] N. Jones. 2018. How to stop data centres from consuming electricity. *Nature* 561, 7722 (2018), 163–166.

[8] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6, 4 (2012), 1–21.

[9] D. Mytton. 2021. Data centre water consumption. *npj Clean Water* 4, 1 (2021), 11.

[10] K. Osei-Bryson and J. Bauer. 2022. Site selection for data centers: A multi-criteria decision analysis. *Journal of Information Technology* 37, 2 (2022), 145–163.