BuLBA-Saar: A Business Location Based Analysis for Saarland

*Research-in-Progress Poster Paper*

Abstract

The goal of this study is to create a tool that can effectively guide new or relocating businesses in Saarland towards a suitable location. We also introduce *Saar-dex*, the first ever comprehensive dataset detailing Saarland’s socio-economic landscape. *Saar-dex* encompasses 28 features, provides a diverse array of information and can be utilized across various tasks and analytical pursuits in the future. The machine learning-powered consulting tool, *BuLBA-Saar*, developed using the *Saar-dex* Dataset has a 0.86 average total of precision, recall and F1-score.

# Introduction

The establishment of a new business requires many significant decisions, among which the selection of a suitable location is crucial. A critical aspect of the decision-making process involves identifying the region that is conducive to the growth and progress of the business.

The literature on economic geography has firmly established that the optimal location of a business yields a number of significant advantages (Liu, Chen, Xu, Chen, & Li, 2021). Yet, to date the socioeconomic profile of the subregions of Saarland with the objective of providing reliable advice to new or relocating companies that seek to (re)locate in the state has not been explored.

To this end, using a multi-layer perceptron classifier and the Gower distance metric (Gower, 1971) we created *BuLBA-Saar* (Business Location Based Analysis for Saarland), a consulting tool that utilizes user-provided information regarding the company to output the most suitable municipality ('Gemeinde') in Saarland. In light of the lack of a preexisting data set, we undertook the task of creating *Saar-dex,* a comprehensive dataset which we believe will prove beneficial for future studies centered on Saarland.

The paper is organized into six sections. The introduction is followed by an overview of related work in the field. The subsequent section outlines the procedure employed, namely the method and metric used to analyze the data and create the model. The fourth section showcases Saar-dex, elucidating the incorporated features and the rationale behind their inclusion. Moving forward, the subsequent sections presents BuLBA-Saar, its structure and the visualization of its predictions.The final section comprises the conclusions, followed by the list of references. Related work

The identification of the optimal location for a new business has been predominantly conducted through the identification of clusters present in a given region. There are two main approaches that differ on their scope, specifically the number of regions that are taken into account.

The multi-region approach creates clusters across different regions, allowing comparative analysis of their economic performance, their composition and the role of clusters (Porter, 2003). Different kinds of cluster definitions have been put forward that differ on how they measure the relatedness with which they perform the clustering process. Different cluster definitions generate different types of clusters (input-output clusters (Feser & Bergmann, 2000), Knowledge-based clusters (Koo, 2005), and co-location based clusters (Porter, 2003)). The aforementioned approaches create a different number of clusters which are either overlapping (Feser & Bergmann, 2000) or mutually exclusive (Porter, 2003). The second approach focuses on identifying linkages between industries and businesses and mapping clusters on a regional level. The comparative advantage of this approach is that it reconciles the inability of multi-region approaches to capture region-specific linkages between industries (Duranton & Overman, 2004).

# Procedure

The study focuses on Saarland, a small post-industrial German state on the border with France. It aims to develop a reliable tool (BuLBA-Saar) that guides companies seeking to (re)locate in Saarland, by recommending the optimal location to found a company based on user-provided data.

The model was trained on *Saar-dex*, a dataset that incorporates a number of features pertaining to the businesses of Saarland and its municipalities. The dataset created for this study is the first of its kind for Saarland, serving as a valuable resource for future research on the region. Nonetheless, *Saar-dex* has apparent limitations. The sparsity of information regarding businesses in Saarland resulted in the exclusion of a significant number of companies from the final data collection.

This proved to be detrimental to the task at hand. In the absence of better alternatives, clustering algorithms group together businesses that are spatially very distant from each other. As a result, this created non-geographically-contiguous clusters that stretched over large parts of Saarland. Moreover, there was significant overlap among these clusters. Consequently, the sparsity of the data renders the cluster creation counterproductive, as the formation of large, spatially dispersed clusters lack informative value and do not provide reliable guidance for new or relocating companies.

Instead, we adapt a radically different approach by tackling the clustering task as a classification task. By doing so, we are able to leverage the available data more effectively. Specifically, we attempt to quantitatively analyze the data in order to be able to reliably classify new companies into the municipality that is the most promising location for them based on the available features.

To achieve that, a multilayer perceptron is employed. Neural networks are trained on numerical data. The features in Saar-dex, however, encompass both numerical and categorical values. In order to transform the values to numerical data in a meaningful way we employ the Gower distance (Gower, 1971), which is designed to handle mixed data types. It calculates the dissimilarity between two data points (e.g., i and j) by averaging the partial dissimilarities across all n features. In numerical features, the partial dissimilarity (PD) between two data points in a given feature is equal to the absolute difference of their values divided by the range of the feature. The range for a given feature is derived by subtracting the lowest from the largest value. For categorical features, the partial dissimilarity is assigned a value of one if the feature values are different, and zero if they are the same.

# Saar-dex

In this section, we introduce *Saar-dex*,the first extensive data set that captures the diverse socio-economic landscape of Saarland. Saar-dex was compiled using data extracted from several public facing online sources. Data sparsity limited the final total number of companies in the dataset to 110. Nevertheless, it enables a detailed exploration and analysis of the region's business landscape, providing valuable insights on the economic dynamics of Saarland.

The final data set amounts to a total of 28 features that can be divided into two distinct categories: business-related and location-related features. The former supply valuable information on the defining characteristics of the companies, while the latter provide a deep understanding of the sub-regions of Saarland. The most important features are presented in the remaining section.

## Business-related features

The data set consists of 110 businesses, each characterized by fourteen business-related features. The first feature of the data set is 'rating' which casts light on the opinions of former and current employees for each organization. The rating values are in the form of interval data ranging from 0.0 to 5.0. Rating of companies is a significant feature as it directly affects the appeal companies have to potential employees. Additionally, rating serves as index of employee retention rates as low rated companies are more likely to suffer from workforce leakages.

The feature 'size' contains a rough approximation of the numbers of employees of the company given as range values ('1 to 50 Employees', '51 to 200 Employees', '201 to 500 Employees', '501 to 1000 Employees', '1001 to 5000 Employees', '5001 to 10000 Employees', '10000+ Employees'). It should be noted that these values refer to the number of employees a company has in total, not just in Saarland. Nevertheless,the feature provides a reasonable estimate for Saarland too, as it is highly unlikely that a company with more than 10,000 employees would have a very small number (e.f. 1 to 50) of employees in Saarland. In fact, in these cases having access to the total number of employees of an organization is even more useful. Companies with large employee bases that have smaller offices in Saarland should be distinguished from actual small and medium-sized enterprises (SMEs) as they have striking differences in terms of resources, scale of operations and economic impact. As with rating, the size of the employee base of the cluster is an index of the existence of a pool of skilled professionals which can be advantageous when recruiting and searching for specialized expertise.

The most important business-related feature is ‘industry’. A threshold of 10 companies was set as a requirement for inclusion of an industry in the data collection in order to ensure that the analysis will have enough data points to be conducted. In total 11 industries surpassed the threshold.

Finally, the data set contains a number of variables (street name, street number, town, postal code, formatted address, latitude, longitude, municipality, municipal key number) that pertain to the location of the business ranging from the very specific (latitude, longitude) to broader spatial and administrative units (town, postal code, municipality, district, municipal key number).

## Location related features

The data set includes a total of fourteen location-related features. These features are crucial at sketching the profile of the various sub-regions within Saarland, their competitive advantages and disadvantages.

'Size of Municipality' indicates in kilometers squared each municipality's size. Furthermore, 5 features describe the population make-up of Saarland (Population, Number of Foreigners per Municipality, Foreigners as % of Population, Raw Population Gain/Loss, Population Gain/Loss Per 1000 Inhabitants). These features should be taken into consideration as the demographics of each region delineate the available labor pool and customer base, both crucial when deciding the location of a business (Peña, 2008; Juyal, 2013).

Furthermore, the data collection features two columns regarding property costs. 'Property prices in € per m²' reflects the cost of real estate. The property tax column (Property tax B in 1000 euros) provides information regarding the costs associated with owning land in each municipality. Higher property prices and taxes have a significant impact on the overall budget and the profitability of the business known as return on investment, while they also undermine employment and sales (Belotti, 2016). Nonetheless, these features also reflect the demand for real estate, investment opportunities, and the overall attractiveness of the area to businesses and residents alike.

Last but not least, to compensate for the exclusion of some businesses due to data sparsity, the presented collection of data contains a feature specifying the industrial sector with most employees per municipality ('Sector with Most Employees'). It is worth bearing in mind that the values of this variable do not perfectly correspond to the 'industry' feature presented in the previous section but rather represent broader definitions of the notion of industry (Agriculture, Forestry and Fisheries, Manufacturing Sector, Trade, Transport and Hospitality,Provision of Business Services, Provision of Public and Private Services). The number of people working in each of these sectors per municipality is additionally included in the dataset.

Although the dataset presented encompasses a wide array of features, it is important to note that not all of them were used for BulBA-Saar. Nevertheless, we have chosen to include these features with the purpose of constructing a comprehensive dataset for Saarland, which holds the potential to facilitate various tasks and analyses in the future.

# BuLBA-Saar: A neural network approach

A total of 17 features (['rating', 'size', 'industry', 'size of municipality', 'population', 'property prices in € per m²', ‘property tax B in 1000 euros’, 'agriculture, forestry and fisheries', 'manufacturing sector', 'trade, transport and hospitality', 'provision of business services', 'provision of public and private services', 'industry with most employees', 'number of foreigners per Municipality', 'foreigners as % of population', 'raw population gain or loss', 'per 1000 Inhabitants'] were utilized for training the model. The classification output is the optimal municipality within Saarland, that is the best choice for the new or relocating business.

The data was split into two parts, training (80%) and testing (20%). This splitting ratio was determined based on the size of the dataset. Given its small size, it was essential that a significant portion of it was used during training. Holding out 20% of the data points for training is a considerable amount, this is in line with common practices in the field (Joseph, 2022).

We utilized the Sklearn library implementation of a multi-layer perceptron classifier with lbfgs as solver and alpha rate of 1e-5 (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel & Duchesnay 2011). The neural network consists of 50 fully connected layers with 24 neurons per layer. The large number of features necessitated the choice of a multi-layer perceptron sacrificing the advantage in explainability provided by other models considered in the design phase.

The test set was used to compute multiple test metrics, namely precision, recall and F1-score. The average total of precision, recall and F1-score is 0.86. It is important to note that, due to the uneven distribution of the dataset among classes, classes with a greater number of companies tend to exhibit higher recall. This means that the model's performance may appear better for the classes that are more prevalent, while relatively poorer for the minority classes.

BuLBA-Saar is accessible through an intuitive, responsive and user-friendly website (<https://saartificial-intelligence-a6bd90a76dd3.herokuapp.com>). The Flask framework was used for website construction along with google maps APIs styling with css and deployed on herokuapp.

In order to determine the optimal location, the user provides relevant company information. The system generates a map that identifies the most suitable municipality for the business, while also highlighting the existing companies that are already established within the municipality.

# Conclusion

The current study developed a consulting tool aimed at assisting new or relocating companies in Saarland. The creation of a comprehensive dataset, encompassing features related to both companies and municipalities, was instrumental in the BuLBA-Saar's development.

A series of different models were tested during the design phase. In the end, a multi-layer perceptron was utilized sacrificing explainability for a higher total average accuracy, recall and F1-score of 0.86. The high evaluation scores verified that the interplay of business and location related features included in the dataset yields better recommendations. The consulting tool is accessible through an interactive, user-friendly website.

All in all, the development of BuLBA-Saar as a consulting tool addresses the critical need for businesses to identify the most suitable location within the state. BuLBA-Saar provides reliable recommendations, empowering businesses to make informed decisions and maximize their chances of success in Saarland's dynamic market environment.

# References

Belotti, F., Porto, E. D., & Santoni, G. 2016. “The effect of local taxes on firm performance: Evidence from geo-referenced data,” *Journal of Regional Science* (61:2), pp. 492-510.

Delgado, M., Porter, M. E., & Stern, S. 2015. “Defining clusters of related industries,” *Journal of Economic Geography* (16:1), pp. 1-38.

Feser, E., & Bergman, E. 2000. “National industry cluster templates: A framework for applied regional cluster analysis,” *Regional Studies* (34), pp. 1-19.

Gower, J. C. 1971. “A general coefficient of similarity and some of its properties,” *Biometrics* (27: 4), pp. 857–871.

Joseph, V. R. 2022. “Optimal ratio for data splitting,” *Statistical Analysis and Data Mining: The ASA Data Science Journal* (15:4), pp. 531–538.

Juyal, S. A. 2013. “Effect of demographic factors on consumer buying behavior of durable goods,” *Indian Journal of Marketing* (43: 12).

Komorowski, M. 2020. “Identifying industry clusters: a critical analysis of the most commonly used methods,” *Regional Studies, Regional Science* (7), pp. 92-100.

Liu, Z., Chen, X., Xu, W., Chen, Y., & Li, X. 2021. “Detecting industry clusters from the bottom up based on co-location patterns mining: A case study in Dongguan, China,” *Environment and Planning B: Urban Analytics and City Science* (48: 9), pp. 2827-2841.

Nielsen, F. 2016. “Hierarchical clustering,” in *Undergraduate Topics in Computer Science*, I. Mackie (eds.), New York: Springer-Verlag, pp. 195-211.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. 2011. “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research* (12), pp.2825–2830.

Peña, I. 2008. *Utility-based data mining: An anthropometric case study*. Retrieved from <https://ruor.uottawa.ca/handle/10393/27723>

Porter, M. 2003. “The economic performance of regions,” *Regional Studies* (37: 6-7), pp. 549-578. Retrieved from <https://doi.org/10.1080/0034340032000108688>

Turban, D. B., & Cable, D. M. 2003. “Firm reputation and applicant pool characteristics,” *Journal of Organizational Behavior* (24:6), pp. 733–751.