Machine Learning based Pricing Methodology for the Logistic Domain: a Preliminary Approach

Antonio L. Amadeu[1], **Fernando Vinturin**[1], **Guilherme A. Zimeo Morais**[1], **Maickel Hubner**[1], **Eduardo M. Pereira**[1], **Marcelo Santos**[1]

[1]Loggi Tecnologia – Data Chapter
Alameda Santos, 2400 – CEP 01418-200 – Sao Paulo – SP – Brazil
R. Jardim do Tabaco 1° Piso – 1100-651 – Lisbon – Portugal

Abstract. *In this work, we introduce a new methodology to discover logistic regions for pricing. We use value-based characteristics from different sources, such as demographic, socioeconomic, risk, transportation, among others, to find homogeneous and valuable pricing regions. The problem was formulated as a traditional cluster solution, where well-know metrics, such as BIC and silhouette score, were used for technical validation, and business premises and constraints, operational and sales, where used to enrich feature engineering and refine cluster formation. The results presented here are from a preliminary work that was validated through several sessions with stakeholders of interest, but it is still missing the market validation. Indeed, this work will be deployed soon and a more detailed validation process, including client adherence, will be performed and monitored until the end of this year.*

1. Introduction

In Logistics domain, goods are charged according to corresponding geographic zone prices. Traditionally, zone prices are defined by several financial and commercial decisions, and by default they are not dynamic and scalable. Such geographic division aims to maximize the opportunity to increase volume and reduce transportation cost, and, consequently, improve margin.

In the present work, we introduce a novel methodology to the characterization of value-based pricing regions. We incorporate multiple proxies, such as socioeconomic potential, risk factors, geographic features, freight prices, among other information, to address the customer-focused pricing perspective, while also focusing on understanding the consumer's perceived value of a transportation service. The expected benefits of this approach are the revenue maximization and the opportunity to improve our transportation network in terms of cost reduction and expansion potential.

A major challenge to accomplish this work was the identification and integration of multiple sources of information to represent a rich data ecosystem to test different business premises and technical hypotheses. We provide herein details about the implemented methodology to address this as well as early obtained results and examples of pricing regions. Our model has been deployed and is currently under validation in real use cases.

2. Background

A pricing region in logistics is a geographic zone that exhibits similar behavior in order to establish a common and coherent pricing. Towards this goal, it is expected to present intra-region coherence and inter-region divergence. These assumptions are the foundation of a machine learning problem known as clustering in an unsupervised learning setting.

We denote cluster analysis as the partitioning of data into meaningful subgroups, whereas the number of subgroups and further information about their composition may be unknown [Hartigan 1975, Fraley and Raftery 1998]. Clustering methods range from largely heuristic to more formal procedures based on statistical models.

In model-based clustering, there is the assumption that data is based on a mixture of underlying probability distributions, in which each component represents a different group or cluster [Fraley and Raftery 1998, Banfield and Raftery 1993]. The Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates, useful in a variety of incomplete-data problems [Gentle 1998, Bock 1996, McLachlan and Chang 2004].

To assess the resulting clusters, quality metrics are usually used. Bayes factors, approximated by the Bayesian information criterion (BIC), assist on determining the number of components and for identifying the partitions most closely related to the input data [Fraley and Raftery 1998, Mateu et al. 2007]. In a more qualitative way, Silhouette score aims to evaluate the data assignment to clusters by measuring the inter-cluster separation and the intra-cluster cohesion [Ogbuabor and Ugwoke 2018, Rousseeuw 1987].

We used Gaussian mixture models (GMM) for clustering and the Silhouette score to assess the quality of the generated clusters. Additionally, k-nearest neighbors (kNN) algorithm is applied to classify instances that did not present enough data to be clustered [Peterson 2009].

3. Methodology

3.1. Data and features

Due to the inner complexity of segmenting well-defined regions that could improve our logistic pricing strategy, we built a rich data ecosystem from diverse data sources as illustrated on Figure 1.

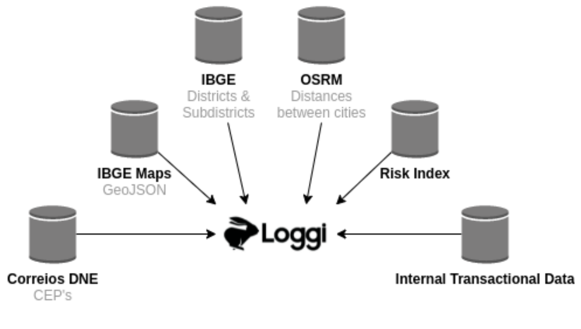

Figure 1. Data ecosystem considering diverse subdomains of interest.

The risk index source has also a predominant importance: it is an in-house dataset that aggregates risk information from Correios, public security information and Loggi's historical information that help to generate more accurate pricing regions. Geographic data from IBGE[1], the Brazilian institute for Geography and Statistics, and Correios[2] are

[1] IBGE shape files for Districts and Subdistricts.

[2] Correios-DNE Brazilian territory with postal code.

useful to obtain a complete list of all Brazilian cities, districts and neighborhoods, their maps and postal codes. The internal transactional data provides information about freight prices, origin, destination, cubic volume, weight, among others. Finally, OSRM[3] is used to generate route information from-to different locations in Brazil.

Based on this data ecosystem, we extracted following features for clustering:

- Distance: estimated distance (OSRM) between pickup and delivery cities;
- Weight: median of packages weight per destination city;
- Risk index: Loggi's loss rate and postal code risk metrics;
- Geographical coordinates: latitude and longitude centroids of city;
- Weight-distance factor: ratio between freight value, weight and distance.

3.2. Clustering

Following some business premises, closely related to our operational team, the clustering is conducted on each federative unit of the Brazilian territory. The idea is to gradually learn and evolve the method, adjusting it on demand to accommodate possible data quality issues or unexpected situations.

Analyzing the data, exclusion of outliers follows and the IQR (interquartile range) was the preferred measure of spread given the data distribution. Additionally, a heuristic is included to remove cities with high variance for freight prices. A customized grid search algorithm for testing a range of hyperparameters following an expectation-maximization (EM) algorithm for fitting the GMM is used. A chart of silhouette scores for different numbers of clusters and covariance types is used to validate results. It is worth mentioning that we used other metrics, such as BIC, however the silhouette score reveals to be the one with closest consensus when compared against business expert decisions. An additional step is considered by applying PCA, retaining only the first two principal components to visually evaluate the quality, in this case separability, of the final clusters.

Finally, cities with few data observations were categorized to the generated clusters through a kNN approach based on geographic coordinates and social-economic indexes as features. As validation for this final step, we generated a visualization based on graph principles for analyzing how the new cities were linked to clusters.

3.3. Validation

To validate the development and results of this work, several business stakeholders were identified and engaged to analyze and discuss the results from an operational and sales point of view. A validation process timeline was designed with sequential sessions to ensure that conclusions are included in the following development step. For instance, the geographic comparison between the generated pricing regions against a calculated internal risk map, with operational specificity, was extremely valuable to validate the clusters proposal from a business perspective.

This process started with an alignment session where relevant concepts and expected functionalities, premises and constraints were discussed, accordingly with different needs. The major conclusions were that a pricing region must be as homogeneous as

[3] OSRM open-source routing machine platform.

possible in terms of price and socioeconomic potential, while being able to obey some operational constraints of Loggi transportation network. It is important to highlight that the main goal of the generated regions is that they can express a value-based price standpoint. In more detail, the development was made in three sessions and two online validation sessions, with several asynchronous validation threads for specific details.

One example of the improvement of these sessions occurred at the first session with stakeholders from Rio de Janeiro State, where the segmentation based on freight median and socioeconomic indexes proved to be insufficient due to some risk factors derived from particular districts. A re-design was made to model and include a compound risk index aiming to obtain a clear alignment with the business expectations. Other variables from our data ecosystem were explored to improve the main intra-convergence and inter-divergence factors and impacts the uniqueness and separability of the clusters from a semantic perspective. Such type of explanations were presented to all the stakeholders.

4. Results

In this section, for the sake of brevity, we provide an example of validation aspect considered for Rio de Janeiro State. Based on the business inputs and alignments obtained from the validation sessions, we were able to capture important feedbacks on the number of clusters per region. Technically, the identification of best K clusters was done with BIC and combining silhouette score for different covariance types. Figure 2 shows such metrics.

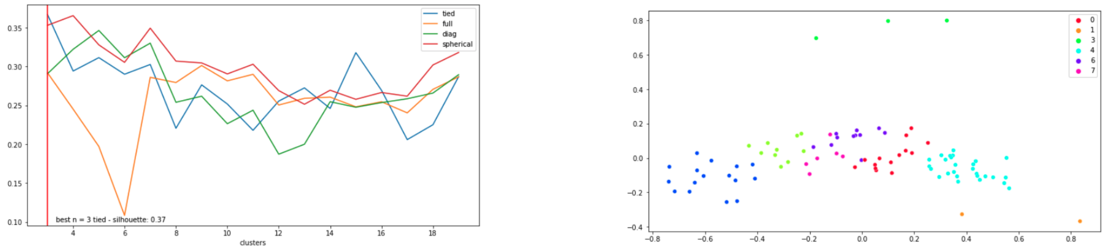

Figure 2. Best *K* considering silhouette score and dimensionality reduction applied to clusters labels for visual validation.

For Rio de Janeiro we started the validation with 3 clusters and finalized the process with 7 clusters. Silhouette analysis helped us to understand the impacts of the number of clusters in terms of clustering quality and it was important in order to accommodate business requirements for specific regions. Also, some discussions implied on the inclusion of other variables to the model. Figure 3 illustrates the results of the clustering process. The left-side map is showing the preliminary set of regions generated and the right-side map the final pricing regions highlighting some improvements after the validation sessions.

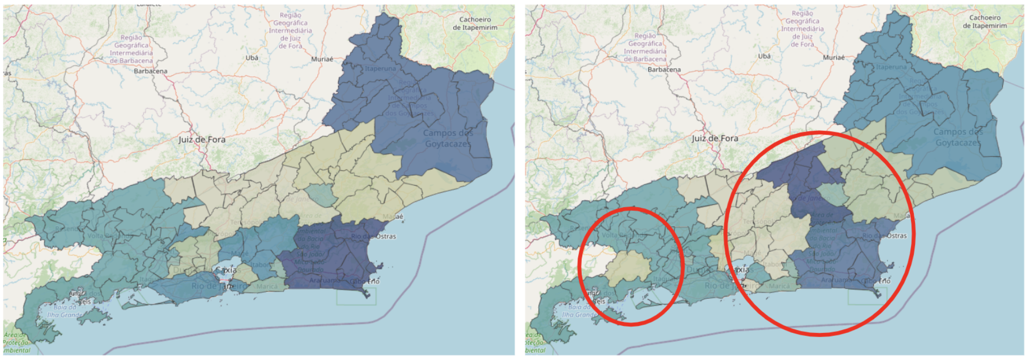
Região Geográfica Intermediária de Barbacena
Região Geográfica Intermediária de Juiz de Fora
Cachoeiro de Itapemirim
Barbacena
Ubá
Muriaé
Juiz de Fora
Campos dos Goytacazes
Região Geográfica Intermediária de Barbacena
Região Geográfica Intermediária de Juiz de Fora
Cachoeiro de Itapemirim
Barbacena
Ubá
Muriaé
Juiz de Fora
Campos dos Goytacazes

Figure 3. Clusters before and after the validation process.

5. Conclusions

This work represents a fresh approach into the traditional way to identify pricing regions for Logistics in the industry. We combine robust and well-studied machine learning algorithms to tackle this problem, while injecting into its formulation and validation several business, mainly operational and sales, constraints and premises. As well, the creation of a rich and diverse data ecosystem along with a consolidated data pipeline made possible to obtain greater stability in modeling the clustering problem, and to optimize the validation process and, consequently, achieve more reliable analyzes.

Clustering-based solutions are largely presented in the literature with many different applications. However, when they are applied to a real business problem within a tight development schedule, the implemented solution normally suffers either from lack of validation or from large validation subjectivity. We tackle this issue by combining technical metrics, such as BIC and silhouette score, and by promoting several business validation sessions with explanatory variables and graphs, whose help us to better tune the GMM parameters, in order to generate clusters with intra-region coherence and inter-region divergence, and to bring added value from business perspective to the process of defining the final pricing regions, respectively.

Defining the pricing regions was a relevant first step towards a dynamic value-based pricing strategy. However, it is still a work in progress initiative. We have deployed this solution and currently we are developing a monitoring framework to measure the impact of these new pricing regions into our pricing model considering different metrics, being one of the most important the adherence of clients. After such validation, we plan to enlarge the data ecosystem and explore more granular clusters in some specific regions of the Brazilian territory.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588.

Gentle, J. (1998). The em algorithm and extensions. *Biometrics*, 54(1):395.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Mateu, J., Lorenzo, G., and Porcu, E. (2007). Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics*, 16(4):968–990.

McLachlan, G. and Chang, S. (2004). Mixture modelling for cluster analysis. *Statistical methods in medical research*, 13(5):347–361.

Ogbuabor, G. and Ugwoke, F. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology*, 10(2):27–37.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.