Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230666

Analyzing Car Configurator Impact Through Genetic Algorithm from a Regional Perspective[1]

Juan Manuel GARCÍA-SÁNCHEZ [a,2], Xavier VILASÍS-CARDONA [a], Álvaro GARCÍA-PIQUER [a] and Alexandre LERMA-MARTÍN [b]

[a] *Research Group of Smart Society, La Salle-Ramon Llull University, 08022 Barcelona, Spain*

[b] *SEAT S.A., 08760 Martorell, Spain*

ORCiD ID: Juan Manuel García-Sánchez https://orcid.org/0000-0003-2569-6561, Xavier Vilasís-Cardona https://orcid.org/0000-0002-1915-9543, Álvaro García-Piquer https://orcid.org/0000-0002-6872-4262

Abstract. This study examine whether visits to the Car Configurator website from a specific area in Spain, referred to as "compound," have a similar impact compared to visits from other locations. The impact is measured by the correlation between clickstream data and sales records. To analyze this relationship, genetic algorithms are employed. The findings reveal that the genetic algorithm surpasses the benchmark values by more than 65 points, indicating its effectiveness. Moreover, the correlation achieved from locations outside the compound is found to be equivalent to the fitness obtained from the regions comprising these compounds. This suggests that the impact of website visits on sales is consistent across different geographic locations.

Keywords. Car Configurator, Genetic Algorithm, Correlation, Automotive Industry

1. Introduction

The Car Configurator (CC) website, offered by car manufacturers, allows customers to select from the company's range of models, estimate a purchase value and schedule a visit to the dealership [4]. This manuscript postulates the possibility of extracting customers' profile from this resource. We pursue to figure out if the visits to the webpage registered from one area of Spain, named compound, have the same impact than the ones received from another location. The impact is measured by to the correlation between clickstream and sales record. Data mining techniques employing genetic algorithms will be utilized. Results show that genetic algorithm outperforms in more than 65 points the benchmark values. Additionally, the correlation achieved from locations outside compound is equivalent to the fitness coming from the regions that make up these compounds.

[1]This work is partially funded by the Department de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2019-34.

[2]Corresponding Author: juanmanuel.g@salle.url.edu

The article is structured in the following way. Firstly, in Section 2, we present related works for the research topic. Hence, Section 3 describes the dataset provided by the automotive OEM source. Next, the methodology and results of the research are in Section 4 and Section 5, respectively. The discussion takes place in Section 6. Finally, Section 7 provides conclusions gained and future research paths.

2. Related Works

In the literature, there is a research study that investigates the optimization of compound delivery time distribution [2]. However, this study did not consider information from customers. Our approach to this task involves extracting consumers' purchasing behavior from the massive and noisy clickstream data. This can be thought of as a feature selection exercise. A comprehensive examination of the current advancements in methods for selecting relevant features can be found in [6]. Genetic algorithms deserve special mention, as they are a type of optimization algorithm. We focus on works that utilize correlation as an assessment metric of the genetic algorithm. This technique has demonstrated its effectiveness in various domains, including: (a) cancer research [5]; (b) build a data warehouse [7]; (c) feature selection in high-dimensional datasets [8]; and (d) identifying apple leaf disease in computer vision [1].

Unfortunately, we were unable to find research that specifically addresses user-generated data from non-transactional webpages. Online activity data can offer non-intrusive means to gain insight into consumers' purchasing behavior.

3. Dataset description

The dataset under analysis spans from April 2017 to January 2020, both included, and pertains exclusively to the Spanish national market. It contains what it has been called car variant, i.e., the user's geographical location (*GeoSeg*) and attributes of the configured car (TRIM, Engine and Color). Additionally, the users' connection day is saved (DOW). The user's geographical location is defined by the Spanish provinces. They are divided by the area of influence of the compound planted by the car manufacturer along the national territory. These compounds act as the company's warehouses. In total, there are six of them, which contain from 4 to 15 provinces each. Additionally, the OEM source has furnished sales records for identical time periods, car models, and geographic regions. Average values from both registers are on Table 1

Table 1. Average values of the weekly volume of configurations contained in clickstream data and sales per compound.

	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**
Avg. [config.]	501.83	8073.13	3249.97	2668.20	7056.11	3212.46
Avg. [sales]	43.94	145.03	130.31	96.07	181.74	126.42

4. Methodology

The genetic algorithm (GA) contains chromosomes made of the attributes defining car variant and DOW the user visited the webpage. The raw clickstream data is filtered using the information in the chromosome, which allows for the creation of CC filtered time

series and to compute the correlation with respect to the sales record, based on findings from note [3]. Nevertheless, we impose a delay of 8-week between both registers. It simulates the client exploration phase and it is based on the automobile manufacturing process. The lagging is executed monthly and results are averaged. If there are no users that meet the chromosome criteria, the fitness is penalized. New chromosomes are created by means of crossover and mutation. In addition, elitism is employed to prevent a decrease in fitness although if it remains constant for consecutive generations, an anti-stagnation systems is triggered.

For each one of the six compounds, GA is executed and the results are compared with baseline values. These baselines are obtained by computing the correlation between sales and online visits made from same compound. Subsequently, a detailed analysis of the best individuals is performed based on how provinces are segmented. Specifically, a fitness comparison is conducted between locations belonging to the compound and those ones out of the area of influence, including Kolmogorov-Smirnov test (KS test) to assess the equivalency between the two distributions.

5. Results

The parameters utilized in the study are: (a) the population size was 300 individuals, (b) each chromosome had 150 distinct rules, (c) there were 200 generations, (d) the tournament probability was set to 0.3 times the population size, (e) the crossover probability was set to 0.9, (f) the mutation probability was set to 1 divided by the population size, and (g) the theoretical maximum fitness value was set to 100.

Table 2 reflects the correlation obtained by the best candidate and the baseline for each of the compound regions. Furthermore, it is examined at province level per compound.

Table 2. Correlation obtained by the best GA solution,the baseline and fitness analysis per location within each compound. *GeoC* stands for provinces that belongs to the compound and *GeoA* signifies locations outside the compound

	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**
Benchmark [R2 Score]	31.08	29.46	30.48	32.08	35.35	29.92
GA Fitness [R2 Score]	96.92	97.22	98.32	98.00	99.25	96.15
Rate GeoC [%]	8.00	14.00	19.33	7.33	20.00	29.33
Rate GeoA [%]	92.00	86.00	80.67	92.67	80.00	70.67
Avg. GeoC	0.34 ±	1.70 ±	2.01 ±	2.24 ±	4.61 ±	1.89 ±
[R2 Score]	0.83	3.29	3.06	2.17	8.53	2.90
Avg. GeoA	2.26 ±	2.48 ±	2.00 ±	2.16 ±	3.00 ±	3.53 ±
[R2 Score]	3.15	4.15	3.45	4.19	4.61	5.72
KS test [p-value]	0.004	0.175	0.336	0.336	0.081	0.983

6. Discussion

In the present analysis from Table 2, all cases exhibited a considerable enhancement in terms of correlation when compared to the baseline values. Among the considered regions, Compound 2 displayed the most significant improvement, while Compound 5

obtained the highest correlation value. On average, the R2 Score was augmented by 66.25 $\pm$ 1.46.

On the other hand, it demonstrates that common locations are underrepresented in all analyzed scenarios. Compound 6 exhibits the highest frequency of common locations. Conversely, Compound 4 has the lowest frequency of common locations. In terms of the individual correlation of each category, KS test proves that fitness distributions are equivalent in each case, but Compound 1. In this last scenario, together with Compound 2 and Compound 6, the external locations provide more fitness that the common ones, as the average fitness is larger. The contrary occurs in Compound 4 and Compound 5. Finally, the difference between the two groups in Compound 3 is not significant.

7. Conclusions

The outcomes given by genetic algorithm outperforms the baseline values for all compounds under analysis. Furthermore, the analysis of the provinces revealed an underrepresentation of locations that define the compound's influence area. Nevertheless, the distribution of individual fitness scores for both groups of locations was found to be equivalent based on KS-test. The reasons behind this behavior cannot be determined with the available information. Therefore, we recommend exploring demographic data, including social and economic factors, as an external source that may shed light on this matter.

To conclude, we recommend that car manufacturers integrate this methodology into their commercial and logistics operations, as genetic algorithms are highly adaptable and customizable to various situations. Further steps involve leveraging online data sources to determine the optimal location for allocating manufactured cars, ensuring they are easily accessible to customers.

References

[1] Chuanlei, Z., Shanwen, Z., Jucheng, Y., Yancui, S., Jia, C.: Apple leaf disease identification using genetic algorithm and correlation based feature selection method. International Journal of Agricultural and Biological Engineering **10**, 74–83 (01 2017).

[2] García-Sánchez, J.M., Cardona, X., Martín, A.: Binary Delivery Time Classification and Vehicle's Reallocation Based on Car Variants. SEAT: A Case Study, pp. 147–150. Frontiers in Artificial Intelligence and Applications, IOS Press (10 2022).

[3] García Sánchez, J.M., Vilasís Cardona, X., Lerma Martín, A.: Influence of Car Configurator Webpage Data from Automotive Manufacturers on Car Sales by Means of Correlation and Forecasting. Forecasting **4**(3), 634–653 (2022). , https://www.mdpi.com/2571-9394/4/3/34

[4] Scholz, M., Dorner, V., Schryen, G., Benlian, A.: A configuration-based recommender system for supporting e-commerce decisions. European Journal of Operational Research **259** (3 2017).

[5] Shah, S., Kusiak, A.: Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine **37**(2), 251–261 (2007). , https://www.sciencedirect.com/science/article/pii/S0010482506000217

[6] Shroff, K.P., Maheta, H.H.: A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy. In: 2015 International Conference on Computer Communication and Informatics (ICCCI). pp. 1–6 (2015).

[7] Tiwari, R., Singh, M.: Correlation-based attribute selection using genetic algorithm. International Journal of Computer Applications **4** (08 2010).

[8] Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 856–863 (2003)