ISBN 978 -987-633-574-4

From Global to Local in the Sneakers Universe: A Data Science Approach

Luciano Perdomo and Leo Ordinez

Laboratorio de Investigación en Informática (LINVI), FI - UNPSJB
Bvd. Brown 3051, Puerto Madryn, Argentina
{lucianor.perdomo,leo.ordinez}@gmail.com

Abstract. In Argentina there was a great growth of e-commerce due to the COVID-19 pandemic. With the aim of helping local companies to understand the market and help them in decision making, data were obtained from online shoe sales sites and with them Machine Learning models were implemented to make price predictions in sneakers. It was concluded that higher-tier companies have greater competitive advantage over lower-tier companies. Nonetheless, the cost-effective methodology used would aid local companies scale up.

Keywords: E-commerce, Machine Learning, Linear Regression, Random Forest, LGBM Regressor

1 Introduction

According to CACE (Argentine Chamber of Electronic Commerce), during 2020, E-commerce turnover in Argentina grew 124% due to the COVID-19 pandemic, compared to the previous year, for a total of ARS 905.143 million, corresponding to more than 164 million purchase orders. The category that grew the most was clothing and sports articles, which in 2019 ranked 4th, and in 2018 3rd [2]. Based on this nation-wide tendency, which replicates also global tendencies towards e-commerce [3,4], an exploratory study was conducted to measure de impact of e-commerce on local stores. In particular, the sector chosen was sneakers (within shoes and clothing) and the territorial scale of the local context is the Patagonian zone in Argentina.

The aim of this research is to build knowledge around products sold through e-commerce channels, which can be leveraged by small local companies, that are joining global tendencies, for decision-making. In particular, through the use of cost-effective tools and information available on the websites of different competitors in distinct scales. This is, the analysis is multiscalar, which is not an impediment considering that e-commerce is inherently horizontal in terms of customers access.

Since the nature of this exploratory analysis is to obtain information publicly available without any intervention inside companies (*e.g.*, asking for sales information) nor action with customers (*e.g.*, surveying preferences), the main variable considered for the products is *price*. In order to predict sneaker prices

ISBN 978 -987-633-574-4

[8], linear regression will be used with price as the dependent variable and gender, brand and company as independent variables. Three experiments were designed. For each experiment different models of Machine Learning [1] are compared and the one with the best results is selected to be optimized and trained. Then, comparisons are made between the models selected above. To make the predictions, efficient models were selected in terms of execution time and resources, and effective in terms of the results.

The rest of the work is organized as follows: in Section 2 the methodology used is outlined; in Section 3 descriptive and inferential results are exposed; a discussion on those results is presented in Section 4; finally, conclusions are drawn in Section 5.

2 Materials and Methods

The standard methodology used in different domains and contexts [9,6] involves understanding the problem; selecting the analytical approach to use depending on the type of research to be carried out at that time; the definition of requirements, the collection and characterization of the data, in an iterative refinement process; the preparation of the data to be able to be worked under the proposed analytical approach, which involves another iterative sub-process of modeling; and the evaluation of the model, which implies its validation by domain experts. After passing the evaluation instance, the model is deployed in an environment available to be accessed. Finally, based on the knowledge obtained, the techniques developed and the products generated in the previous steps, the goal is to obtain learning that promotes better decision-making.

As previously mentioned, we are interested in analyzing Patagonian online sneakers stores in the context of bigger scales, such as nation-wide or globally. Depending on the scope of the company that owns the e-commerce site, it was decided to categorize them as local, national or regional and global. For this, the cities where the stores are physically located and the number of branch offices were considered, where it applies. This is, in the first place, we consider companies which have a physical store; and secondly, in some cases the number of branch offices was not taken as a limit for categorizing but an indicator, and the reach of their marketing strategy was considered (*i.e.*, advertising in international sports events). Seven sites were selected out of seventeen. Globally, Stockcenter, a subsidiary of NetShoes, was chosen. At the National level, Dash and Solodeportes were selected. At a regional level, Sporting. Finally, at the local level, Ferreira (Bahía Blanca and South West of Buenos Aires Province), Quonam (Chubut, Patagonia) and Newsport (Córdoba) were considered. Each company has different types of shoe offerings (Men, Women, Unisex, Children, Boy/Girl). For simplicity and homogeneity of data, it was decided to analyze offerings by categories “Women” and “Men”.

The scraping tools used were parseHub (for Quonam and Stockcenter) and the Python library Beautiful Soup for the rest of the sites. The following data

ISBN 978 -987-633-574-4

were obtained *brand*, *model*, *list price* (price without discount), *net price* (price with discount) and *sex*.

The dataset was cleaned and structured as follows: *brand*, represents the brand of the sneaker; *footwear*, represents the type of shoe; *sex*, women or men; *original_model*, text from the original dataset, that is, without parsing; *model*, parsed text; *net_price*, discounted price applied; *list_price*, price without discount applied; *item discount*, percentage of discount applied; *company*, name of the company where the data was extracted.

As said before, the tools used were ParseHub, Python3, and the following Python packages: Jupyter-Notebook, Pandas, BeautifulSoup, ScikitLearn, Seaborn, plotly-express, XGBoost, LightGBM, yellowbrick, hyperopt.

3 Results

The scraping was carried out on March 19, 2021. Then the data set was cleaned up and structured. The exploratory data analysis was performed. Then the predictive analysis was performed, in which three experiments were carried out to predict sneakers prices.

3.1 Exploratory Data Analysis

In the first place, we considered the variables list price, brands, sex and company. A comparison among them is presented in Fig. 1.

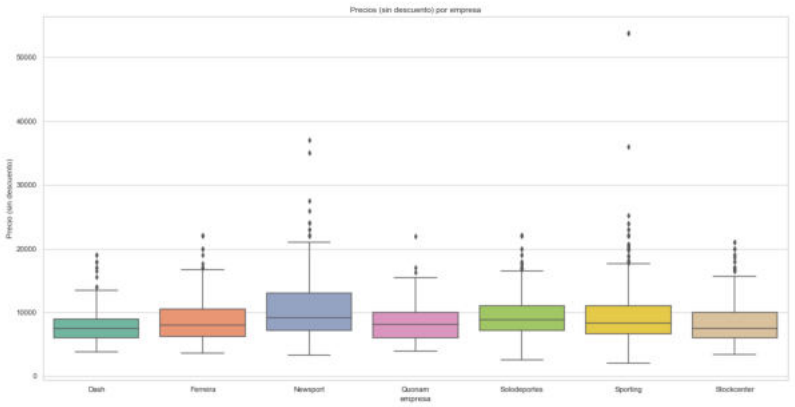

Fig. 1. List prices per company.

Fig. 2 shows the amount of sneakers that each company offers by sex.

ISBN 978 -987-633-574-4

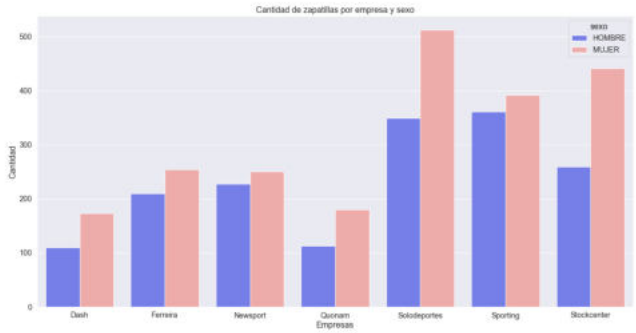

Fig. 2. Amount of sneakers by company and sex.

The amount of sneakers with prices ranging from $2.000 to $30.000 is presented in Fig. 3.

The distribution of prices within the companies is presented in Fig. 4.

Finally, a comparison of characteristics among each company is performed by a radar chart and presented in Fig. 5. The characteristics are as follows:

- Variables: Maximum price, quantity of men's sneakers, quantity of women's sneakers, number of brands, brand dispersion (represented as "HHI Marcas").
- For the dispersion of marks, the Blau Index[10] was used, which quantifies the probability that two individuals taken at random from a population are in different categories of one variable.
- The data was scaled to be in an approximate range of 100 to 1.000, for visualization purposes.
  - The max prices were divided by 100.
  - The Blau index was multiplied by 1,000.
  - The number of brands multiplied by 10.
- With all these data, a Radar Chart was made for each company. Then an overlay radar chart was made to compare all companies.

3.2 Predictive Analysis

In the first place, outliers were removed from the data set, leaving a maximum price of $25.000 and a minimum of $4.000. In addition, brands that have less than 40 items were removed, resulting in 92.09% of the data set. With this, 75% of the data was used for training. Training and test data subsets were saved with the Python library Pickle.

In all cases, it is used as a dependent variable the list price and as independent variables, in turn, brand, company and sex. The following experiments were conducted:

ISBN 978 -987-633-574-4

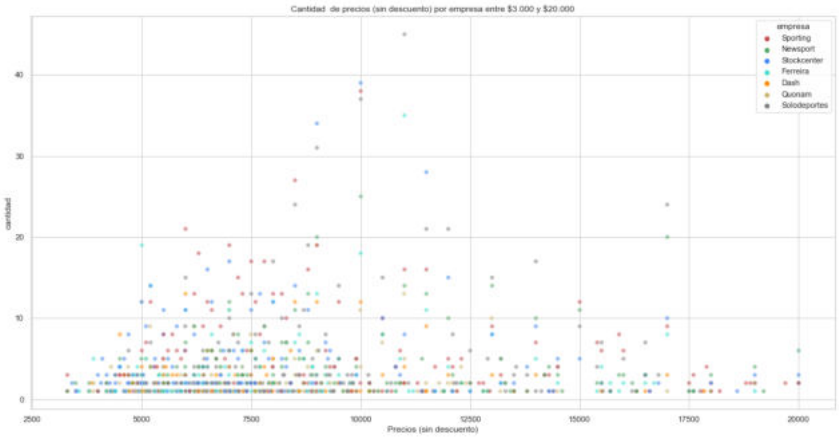

Fig. 3. Amount of prices per company.

1. Comparison of linear models: Linear Regression, Ridge Regression and SGD Regression.
2. Comparison of: Random Forest, XGBoost and Decision Tree Regressor.
3. Comparison of: SVM Regressor, Random Forest and Light GBM Regressor.

Finally, a comparison of the experiments was carried out.

Experiment 1 At first, the Recursive Feature Elimination (RFE) model was used to obtain the number of optimal features, but the idea was discarded, and it was decided to train the model with 3 features, then 2 and finally 1, and then make comparisons. The models were found to better fit a polynomial function.

Ridge Regression It is similar to linear regression, but uses L2 regularization. Hyperparameters can be adjusted to find the correct alpha value, which is the parameter with which you can make the model perform overfitting or underfitting.

The alpha parameter is searched with the yellowbrick library and a value of 1.6907141034735782 was obtained. It was also found that a polynomial function of degree 11 fits the model better, to create a Polynomial Ridge. Iterating between the three features, the best r2 that was obtained was 33.32% with 2 features (brand, sex).

SGD (Stochastic Gradient Descent) Regression It is a linear model that uses L2 regularization and minimize empirical loss with SGD (loss gradient is estimated for each sample and the model is updated with the learning rate). It is better suited to linear models than Ridge Regression and Linear Regression.

ISBN 978 -987-633-574-4

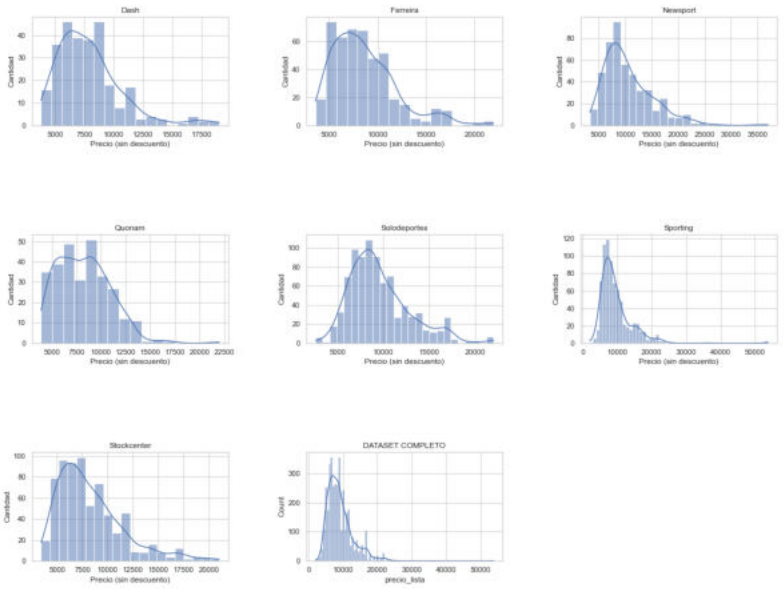

Fig. 4. Distribution of prices.

With the default model and data, a r2 of 2% was obtained. Finally, it was iterated with the 3 features, then with 2 and last with 1, and the polynomial function with 14 degrees. The best r2 result was 6.28% with 2 features.

Lineal Regression The model was trained with a polynomial function of degree 10, giving the following results:

- 1 feature (brand): r2 = 49.47%
- 2 features (brand, sex): r2 = 49.66%
- 3 features (brand, sex, company): r2 = 50.46%

Based on this, it was decided to train each company with a different model, taking into account that the results of linear regression are better than the other two models (Ridge and SGD);

One model was made per company, with polynomial degree 10 and using brand and sex as features.

Following, companies and best r2 (with one or two features) are shown:

- Dash: 1 feature r2=59.11% MAE=0.0600 MSE=0.0082
- Ferreira: 1 feature r2=24.41% MAE=0.0973 MSE=0.0228
- Newsport: 2 features r2=37.1% MAE=0.1331 MSE=0.0317

ISBN 978 -987-633-574-4

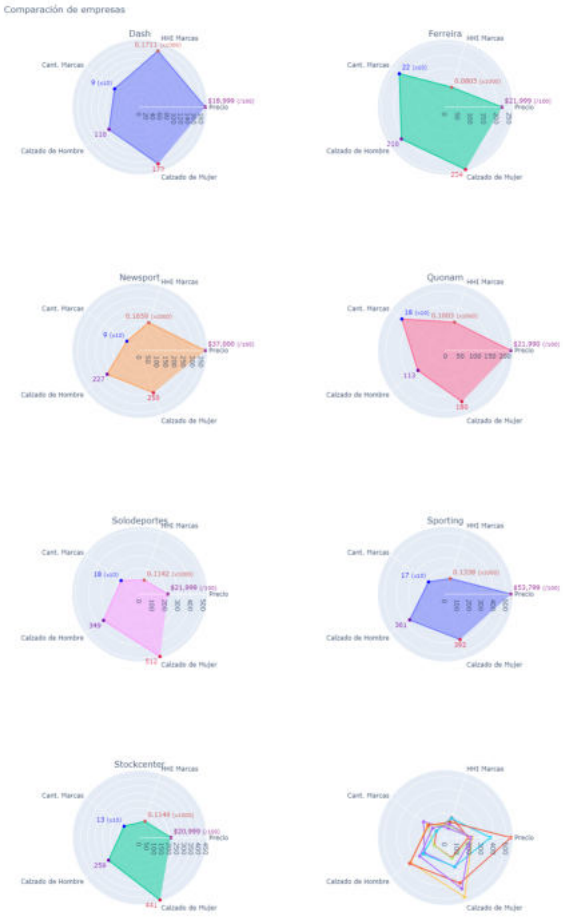

Fig. 5. Radar Chart with all companies.

- Quonam: 1 feature r2=53.66% MAE=0.0634 MSE=0.0068
- Solodeportes: 1 feature r2=33.03% MAE=0.0937 MSE=0.0140
- Sporting: 1 feature r2=44.32% MAE=0.1006 MSE=0.0187
- Stockcenter: 1 feature r2=68.35% MAE=0.0635 MSE=0.0069

Experiment 2 This experiment is based on [7]. Here, Random Forest, XGBoost and Decision Tree Regressor were chosen. Random Forest because each tree draws a different sample, avoiding overfitting and improving the accuracy of predictions. XGBoost (Extreme Gradient Boosting) which, as a Gradient Boosting algorithm, generalize Boosting models as I/O to get better models (they are trained sequentially). Finally Decision Tree Regressor is used as a comparison against the previous two, besides being simple and effective.

ISBN 978 -987-633-574-4

The models were compared, each one with its default parameters. The results were:

- Decision Tree Regressor r2 = 50.20%
- XGBoost r2 = 50.5204% (chosen model)
- Random Forest r2 = 50.5157% (secondary model)

XGBoost Regressor First we tried using the polynomial function with 3 and 6 degrees, but in no case was an r2 of 43.7% exceeded, a value lower than the r2 of 50.52% of the previous comparison, so the hyperopt library was used for parameter optimization, along with three sets of different parameters to test the model and performing tests using one, two and three features: (brand), (brand, sex) and (brand, sex, company).

The best result of all the tests was r2 of 50.32%, with three features, no polynomial function. Since the optimization did not work better than the default parameters, it was decided to try Random Forest, which in the initial comparison yielded similar values.

Random Forest Regression Since there was little difference, it was decided to use it in the experiment. With the default parameters, an R2 = 50.49% was obtained. The search for hyperparameters was carried out with hyperopt. Four different parameters were used; without polynomial function, and with polynomial function of 3 and 6 degrees; with one, two and three features; and parameter max_evals=100 (hyperopt). The best r2 was 51,237%; with 3 features, polynomial function of degree 6. Because it gives better results than XGBoost, Random Forest is used for the iteration of each company, with the parameters that hyperopt showed in the winning test.

Companies and best r2 (with one or two features):

- Dash: 1 feature r2=58.85% MAE=0.0603 MSE=0.0083
- Ferreira: 1 feature r2=25.7% MAE=0.0948 MSE=0.0224
- Newsport: 2 features r2=37.74% MAE=0.1301 MSE=0.0314
- Quonam: 1 feature r2=53.85% MAE=0.0629 MSE=0.0068
- Solodeportes: 1 feature r2=34.56% MAE=0.092 MSE=0.01368
- Sporting: 1 feature r2=44.09% MAE=0.1011 MSE=0.0188
- Stockcenter: 1 feature r2=68.33% MAE=0.0632 MSE=0.0069

Experiment 3 It is based on [5]. Here, SVM, Random Forest and LightGBM Regressor were chosen. SVM is more accurate than Linear Regression and by default it uses a linear RGB kernel. Random Forest, for the same reasons as the previous experiment it is used as an indicator within this experiment, due to its use in the previous experiment. LightGBM Regressor, is a Gradient Boosting model, similar to XGBoost, uses algorithms based on decision trees.

The models were compared with the default parameters:

- SVM r2 = 0.1638
- Random Forest r2 = 0.5046 (secondary model)
- LGBM r2= 0.5084 (chosen model)

ISBN 978 -987-633-574-4

LightGBM Regressor For parameter optimization, first, we tried to obtain r2 with the polynomial function of degree 3 (r2=50.74%) and degree 5 (r2=50.67%). The previous comparison (without polynomial function) gives a better result than XGBoost in the previous experiment. Hyperopt was used to find the best hyperparameters, three sets of different parameters, with one, two and three features, and the variable max_evals=100 were used. The best result was with 3 features and without a polynomial function. Because the model yields a promising r2 value, it is used to make predictions for each company and then compare results.

Companies and best r2 (with one or two features) with default parameters:

- Dash: 1 feature r2=59.17% MAE=0.0606 MSE=0.0082
- Ferreira: 1 feature r2=22.29% MAE=0.1001 MSE=0.0234
- Newsport: 2 features r2=34.31% MAE=0.1357 MSE=0.0332
- Quonam: 1 feature r2=15.18% MAE=0.0899 MSE(0.0125)
- Solodeportes: 1f r2=32.64% MAE=0.0944 MSE=0.0140
- Sporting: 1 feature r2=42.11% MAE=0.1021 MSE=0.0194
- Stockcenter: 1 feature r2=64.03% MAE=0.0659 MSE=0.0078

Companies and best r2 (with three features) using optimized parameters:

- Dash: 2 features r2=59.24% MAE=0.0606 MSE=0.0082
- Ferreira: 1 feature r2=22.75% MAE=0.0997 MSE=0.0233
- Newsport: 2 features r2=34.98% MAE=0.1349 MSE=0.0328
- Quonam: 1 features r2=15.28% MAE=0.0899 MSE=0.0125
- Solodeportes: 1 feature r2=33.34% MAE=0.0936 MSE=0.0139
- Sporting: 2 features r2=42.12% MAE=0.1021 MSE=0.0194
- Stockcenter: 1 feature r2=68.35% MAE=0.0635 MSE=0.0069

4 Discussion

It can be seen in the comparison of the companies, that the determination coefficient in the Stockcenter predictions (global category), is the most accurate of all.

Dash and Sporting companies have National or Regional category. Dash is in second place and Sporting in the last experiment takes third place and fourth place in the first two.

The companies at the local level are Ferreira, Quonam and Newsport. Quonam is third in the first two experiments and last in the third one. The companies Newsport, Solodeportes (regional) and Ferreira maintain their order in all the experiments, being fifth, sixth and seventh in the first two and gaining a position in the last.

Except for Quonam company, in the first two experiments, it is true that the higher order companies have better price predictions. This company may have been benefited from cleaning out the outliers in the dataset when training the models.

ISBN 978 -987-633-574-4

On the other hand, the number of brands offered by each company and the number of sneakers per brand, allow companies to compete in different market segments. However, as shown in the comparison of prices of the exploratory analysis, there are no big differences in the dispersion of them. This may be because certain brands impose to be narrowed to certain price ranges.

5 Conclusions

In this work, a data science approach was performed over a local market sector in the context of bigger scale competitors. Although the case study was specific, such as online sale of sneakers, the methodology used was proven to be cost-effective and adaptable to other situations. With that, an analysis of a small particular company can be performed.

Including sales data to the analysis would allow the definition of better marketing strategies. Thus allowing local companies to start competing with national companies; and the national ones with the global ones.

On the other hand, the work in this paper would allow the sneaker buyer to make a more sound decision when buying, not lead by advertisements and publicity.

As future work, it is possible to scrape data weekly from each of the e-commerce sites to create a data warehouse and also collecting more information about the sneakers (such as color, sizes, etc.). Insights can be obtained, for example on the brand/price relationship or between brands.

References

1. B. Boehmke and B. Greenwell. *Hands-on machine learning with R.* Chapman and Hall/CRC, 2019.
2. CACE. Estudio anual de comercio electrónico 2020.
3. W. E. Forum. Covid-19 has reshaped last-mile logistics, with e-commerce deliveries rising 25
4. M. Keenan. Global ecommerce explained: Stats and trends to watch in 2021, 05 2021.
5. A. Kumar. Price prediction using machine learning regression — a case study.
6. I. Martinez, E. Viles, and I. G. Olaizola. Data science methodologies: Current challenges and future approaches. *Big Data Research*, 24:100183, 2021.
7. L. Norman. Predicting stockx sneaker prices with machine learning.
8. D. Raditya, N. E. P, F. A. S, and N. Hanafiah. Predicting sneaker resale prices using machine learning. *Procedia Computer Science*, 179:533–540, 2021. 5th International Conference on Computer Science and Computational Intelligence 2020.
9. J. B. Rollins. Foundational methodology for data science. IBM Analytics, 2015.
10. A. Solanas, R. Selvam, J. Navarro, and D. Leiva. Some common indices of group diversity: Upper boundaries. *Psychological reports*, 111:777–96, 12 2012.