Machine Learning in the Prediction of Net Sales for Colombian Companies in a Post-pandemic Scenario.

Rubén Darío Acosta-Velásquez, William Stive Fajardo-Moreno and Leonardo Espinosa-Leal.

Abstract— The uncertainty about how the economic reactivation will behave worldwide is a general concern; in the face of this panorama, it is essential to look for historical data that allow us to build the present and predict the future, with this purpose and taking advantage of the advancement of technology in the field of Machine Learning, the present work established the predictions on net sales by companies operating in Colombia. To this research, about two million official records were used from the open data portal of the Bogotá Chamber of Commerce, which were divided 70% for training and 30% for tests; based on these data, Linear Regression algorithms were used (LR), Random Forest (RF), XGBoost (XGB), and Extreme Learning Machine (ELM) to make predictions. The results of the regression performance were evaluated through the coefficient of determination, and the best measure performance showed 0,9 with a Random Forest regressor (RF) algorithm.

Keywords—Machine Learning, Prediction, Net sales, Colombia, Companies.

I. Introduction

The current developments and technologies in information storage and data analysis have increased the power of computing. With this, the possibility of running more sophisticated algorithms, for instance, machine learning models that improve the prediction capacity day by day. These tools were some of the main allies to process the data produced in the different scenarios occasioned for the covid 19 pandemic. Endured most of the economic crisis until the bankruptcy, evidenced in the net sales earned during this period.

In this work, we study and present results concerning the prediction of net sales registered in the chamber of commerce of Bogota, Colombia, during the period 2016 until 2019. This data is updated every year. Hence, one of the features recorded is net sales. The original information contains other valuable information such as size, legal type, economic sector, number of employees. This data can be used for analytics purposes using cutting-edge statistical and artificial intelligence methods. Thus, we model net sales companies considering those with net sales are different from zero. Upon a full statistical description of the dataset, where each variable is presented and described, a feature selection process of the data is performed. In the final stage, we split the dataset into two subsets, one for training and the other for testing, then fit four different machine learning methods: Linear Regression algorithms (LR), Random Forest Regressor (RF), XGBoost Regressor (XGB), and Extreme Learning Machine (ELM) to make predictions.

II. Research Settings

The data considered in this research was taken from the Bogota Chamber of Commerce's open data portal, a non-profit centre that promotes business development in Colombia and encourages the constitution of new undertakings. As part of this activities, it is the management of data and statistics. This information is available at https://opendatabogota.ccb.org.co/.

To know the data, the features identified in this analysis are in Table I; in addition, due to different features, pre-

Rubén Darío Acosta Velásquez
EAN University
Colombia

William Stive Fajardo-Moreno
EAN University
Colombia

Leonardo Espinosa-Leal
Arcada University of Applied Sciences
Finland

https://doi.org/10.17758/EIRAI11.F1221104

processing work was necessary; for instance, some of them are specified as dummies.

Furthermore, the feature target is a quantitative value; thus, the models ran were regressors. Their performance was evaluated through the coefficient of determination whose closeness to 1 depict a better performance and the Mean Squared Error (MSE), whose the lowest values show a better performance.

We have used four different machine learning models to predict the disappearance: Logistic regression (LG), Random Forest (RF), extreme Gradient Boosting (XGB) and Extreme Learning Machine (ELM), a single-layer feed-forward Network (Huang, 2006) with proved capabilities in many areas of research such as computer vision (Espinosa-Leal, 2019), time-series, clustering, edge devices (Akusok, 2019a). The first three models were implemented using the Python programming language through the widely used scikit-learn implementation (Pedregosa, 2011). The latter was implemented using a new python library named scikit-elm (Akusok, 2019b). Both random forest models were run using the Python Dask library for parallelization (Rocklin, 2015). We have run a randomized grid search with five-fold cross validation for the models to obtain the best parameters in order to get better values of performance.

III. Descriptive Analysis

The data set used for this work formerly was a data set of around 3 million records, but after a pre-processing taking into account those records different to 0, remain 261972 records in the data set. These registers were omitted because of depicting around 90% of the data set, and the main interest is companies with profits.

Table II shows some descriptive statistics about the feature target scaled and not scaled per year.

Besides, the distribution for legal organization shows that around 80% of companies are considered as a natural person or simplified joint-stock company Fig. 1. On the other hand, the size of the companies the 80% of them are focused on microenterprises Fig. 2, moreover, the number of companies with profits between 2017 and 2020, is according to years before and during covid 19 pandemic Fig. 3, where the decrease in 2020 coincides with the closure of business.

TABLE I.
FEATURES USED IN MODELS

Variable	Type	Values	Variable dummy
Commercial register renewal	Binary	1-Renovated 0-Canceled	No
Legal Organization	Qualitative	1-Anonymous 2-Collective 3-Limited by shares 4-Simple Comandita 5-Associative work company 6-State industrial and commercial company 7-Sole proprietorship 8-Foreigner 9-Limited 10-Natural person 11-Simplified joint stock company 12-Retirement pension fund	Yes
Size	Qualitative	1-Large 2-Median 3-Microenterprises 4-Small	Yes
Sector	Qualitative	1-Agricultural 2-Trade 3-Construction 4-Industry 5-Mines and quarries 6-Services	Yes
Number of establishments	Quantitative	Values between 1 and 2297	No
Number of employed personnel	Quantitative	Values between 1 and 90,000	No
Importer / Exporter	Qualitative	0-No data 1-Exporter 2-Importer / Exporter 3-Importer	Yes
Total asset value	Quantitative	Values between 1 and 1.91242E + 15	No
Total value of liabilities	Quantitative	Valores entre 1 y 3,13034E+15	No
Total equity value	Quantitative	Values between 1 and 1.91242E + 15	No
Total net sales	Quantitative	Values between 1 and 9.06101E + 14	No

TABLE II.
DESCRIPTIVE PARAMETER FOR SCALED AND NON-SCALED NET SALES

Parameter		**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
2017	***Non Scaled***	62646	3,23e10	5,12e12	1	1,47e7	6,22e7	3,59e8	9,06e14
	Scaled	62646	18,05	2,65	0,69	16,5	17,94	19,7	34,44
2018	***Non Scaled***	64636	3,13e10	5,04e12	1	1,5e7	6,31e7	3,61e8	9,06e14
	Scaled	64636	18,06	2,64	0,69	16,52	17,96	19,7	34,44
2019	***Non Scaled***	84477	4,46e10	6,23e12	1	1,26e7	5e7	2,2e8	9,06e14
	Scaled	84477	17,71	2,55	0,69	16,34	17,72	19,21	34,44
2020	***Non Scaled***	50213	5,63e10	7e12	1	1,54e7	6,53e7	3,03e8	9,06e14
	Scaled	50213	17,91	2,71	0,69	16,55	17,99	19,53	34,44

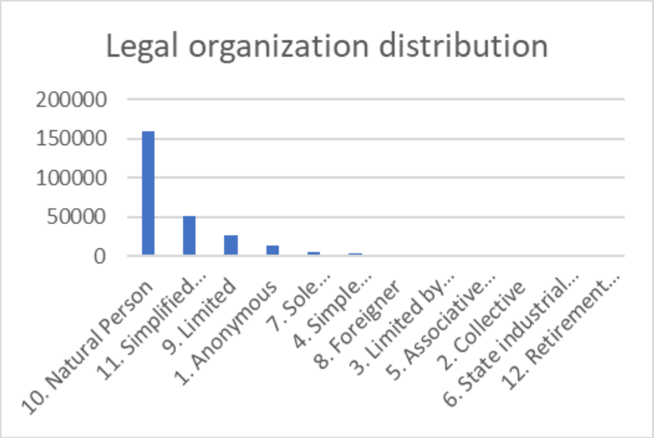

Fig. 1 Distribution of legal organization

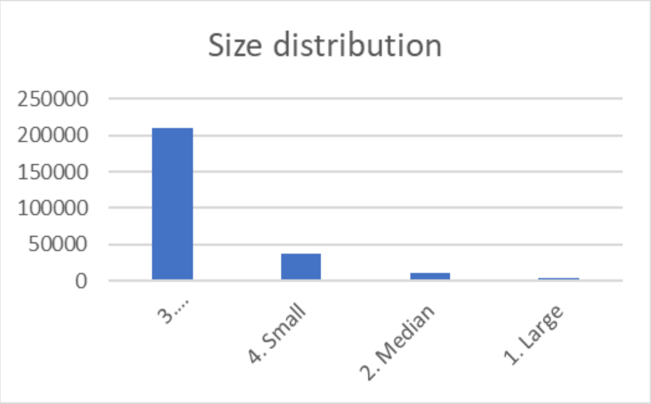

Fig. 2 Distribution of size of the companies

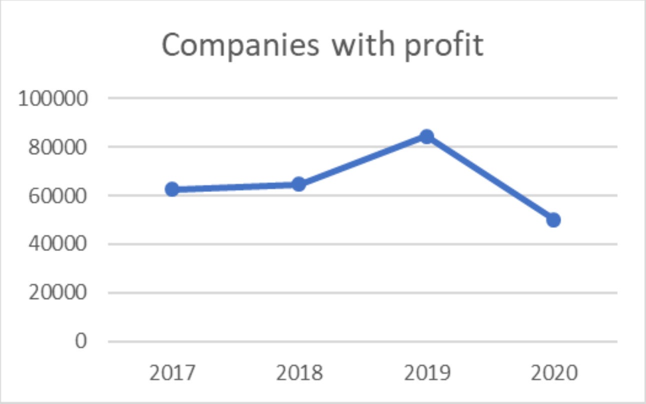

Fig. 3 Number of companies with profits

IV. Modelling

A feature of interest to this study is the net sales of the Colombian companies given the aftermath caused by the covid 19 pandemic so that some regression models as Linear Regression (LR), Random Forest Regressor (RF), XGBoost Regressor (XGB), and Extreme Learning Machine (ELM) were run to make predictions There was a randomized grid search to obtain the best parameters to get better performance values in all cases.

Due to the frequency distribution of quantitative features even the net sales, some machine learning models were run, but the performance measures obtained were poor, for instance, coefficients of determination of 0 and the better of cases 0.77 for a Randon Forest Regressor and though it could be considered as a good performance measure given the obtained with other models, for the above, a logarithm transformation was used to re-scaled the quantitative features with that the performance of the models enhanced. Thus, the models were reset with the features transformed and run to, which improved the performance measures where the random forest regressor showed the best performance.

V.Findings

As we present above, four regression models were run where the feature target is net sales of companies in Colombia before and during pandemic scenarios; in Fig. 4, we present the performance measures obtained in each case.

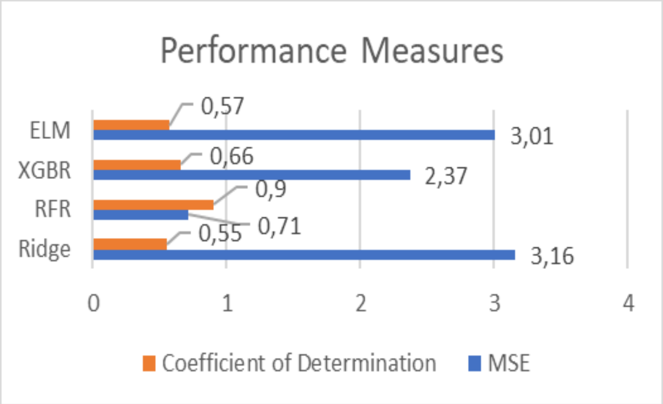

Fig. 4 Comparative of performance measures for each model

According to the results obtained with the previous models with the features without transform, random forest shows the best performance and XGBR as the second-best. However, XGBR presents overfitting, and the other models showed a coefficient of determination of 0.

On the other hand, the performance measures obtained considering the log-scaled target feature presented better results in all models. However, Random Forest remains the model with the best performance with the less MSE (0.71) and the best coefficient of determination closest to 1 (0.9).

VI. Conclusion

The covid 19 pandemic depict a milestone in human history not just because human lives were lost, but also for the damages caused at all levels as an economic level; many companies around the world were closed due to the sudden paralyzation of the world economy and with this the decreasing of their incomes. For the above, machine learning algorithms allow us build models to predict different features of interest and in future events anticipate adverse scenarios.

Acknowledgment

The authors thank EAN University for the support received in the development of the research, as well as the funding for participation in ICISET 21.

References

[1] Akusok, A., Leal, L. E., Björk, K. M., & Lendasse, A. (2019, December). High-performance ELM for memory constrained edge computing devices with metal performance shaders. In *International Conference on Extreme Learning Machine* (pp. 79-88). Springer, Cham. https://doi.org/10.1007/978-3-030-58989-9_9

[2] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, *70*(1-3), pp. 489-501. https://doi.org/10.1016/j.neucom.2005.12.126

[3] Akusok, A., Leal, L. E., Björk, K. M., & Lendasse, A. (2019, December). Scikit-ELM: an extreme learning machine toolbox for dynamic and scalable learning. In *International Conference on Extreme Learning Machine* (pp. 69-78). Springer, Cham. https://doi.org/10.1007/978-3-030-58989-9_8

[4] Espinosa-Leal, L., Akusok, A., Lendasse, A., & Björk, K. M. (2019, December). Website Classification from Webpage Renders. In *International Conference on Extreme Learning Machine* (pp. 41-50). Springer, Cham. https://doi.org/10.1007/978-3-030-58989-9_5

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[6] Rocklin, M. (2015, July). Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference* (Vol. 126). Austin, TX: SciPy. https://doi.org/10.25080/Majora-7b98e3ed-013

[7] Fajardo-Moreno, W. S. (2020). Prediction of the Disappearance of Companies From the Market in Bogotá, Colombia Using Machine Learning. In *Handbook of Research on Management Techniques and Sustainability Disruptive Situations in Corporate Settings* (pp. 227-246). IGI Global. https://doi.org/10.4018/978-1-7998-8185-8.ch011

[8] Acosta-González, E., & Fernández-Rodríguez, F. (2008). Predicción del Fracaso Empresarial Mediante el Uso de Algorítmos Genéticos. Las Palmas de Gran Canaria.

[9] Alaminos, D., Del Castillo, A., & Fernandez, M. A. (2016). A Global Model for Bankruptcy Prediction. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0166693

[10] Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, pp. 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[11] Espinosa-Leal, L., Chapman, A., & Westerlund, M. (2020). Autonomous Industrial Management via Reinforcement Learning Towards Self-Learning Agents for Decision-Making. *Journal of Intelligent & Fuzzy Systems*, *39*(6), 8427-8439. https://doi.org/10.3233/JIFS-189161

[12] Jabeur, S. B. (2017). Bankruptcy prediction using Partial Least Squares Logistic Regression. *Journal of Retailing and Consumer Services*, pp. 197-202. https://doi.org/10.1016/j.jretconser.2017.02.005

[13] Montero-Casarejos, Á. (2016). *Predicción de Quiebras Empresariales Mediante Ingeligencia Artificial.* Madrid. Spain.

[14] Zoricak, M., Gnip, P., Drotar, P., & Gazda, V. (2020). Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. Economic Modelling, pp. 165-176. https://doi.org/10.1016/j.econmod.2019.04.003

[15] Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *162*, pp. 895-899. https://doi.org/10.1016/j.procs.2019.12.065

About Author (s):

Ruben Dario Acosta-Velasquez

Ruben Dario Acosta-Velasquez works as associate professor at EAN University (COL), he is a mathematician, specialist in Operation Research and MSc degree in Applied Mathematics, his research interests are machine learning, deep learning and overall in AI areas. Currently he is the director of Basic Sciences school in EAN University.