THE ABSENCE RESOURCES THAT CAN BE RECOVERED: A proactive approach to an old question with intensive use of artificial intelligence

Tomelin,H.H.S.. htomelin@outlook.com, Oliveira,R.A.M.. ricardomello3355@gmail.com, Oliveira Jr,S.M.. sidney.marques.mp@gmail.com, Zschoronack,T.. thiago.zschornack@gmail.com, Maia,R.G.B.. rafamp@gmail.com

Highlights:

- Reduction in losses due to defaults.
- conscious action versus instinctive reaction.
- Preventive detection of defaulters by recent punctuality in payments.

Keywords: AutoML; Sanitation; Default

INTRODUCTION

Sanitation is fundamental to human life (Sinharoy et al., 2019). Brazil's sanitation services are tariffed (Borges et al., 2022). When a user does not pay the tariff, the others must pay for the defaulter. Actual procedures are ordinary, low effectiveness with high stress level between users and service providers. The specific prediction of the defaults enables preventive actions to reduce the consequences of stress.

To develop a prediction model, data from a city with good levels of services, Joinville, SC, was used. Looking at the CAJ data, we can see that: The default amounts, over four years, are equivalent to the amount needed for a sewage collection and disposal network that would serve 30,000 people, five per cent of Joinville's population.

Currently, there are 2 actions in these cases: notification of defaulter and disconnection of the water supply (delay greater than 50 da). It is a reactive action, with an additional cost of 4 per cent of revenues, time-consuming and inefficient, placing a double burden on defaulters (CUI, 2022). Tension is created between users and the company, which can lead to unbearable stress. By adopting a strategy of active and preventive action, rather than merely reactive action, it is necessary to predict who the defaulter will be before they become a defaulter

The analysis was based on the evolution of payment terms over 24 months, using AutoML tools powered by Knime analytics platform[1]. The result is a prediction model that can determine whether or not the user will enter the default condition.

METODOLOGHY

According to Varian (1996) and Ferguson (1999), default is defined as late payment by the consumer (disengagement) or even non-payment. This generates a cost that is seen as systemic (independent of the organisation's internal management). Cui (2021) reinforces this concept and highlights the financial and institutional stress that is generated in organisations.

The analysis of consumer behaviour in situations of default has been widely studied concerning its overall economic impact and mitigation strategies. According to Shaik et al. (2020), default can lead to a significant increase in the operational costs of institutions, as companies need to leverage financial reserves to cover deficits caused by unpaid bills. Additionally, Kiss et al. (2018) highlight that the credit risk associated with default can affect the financial stability of an organization over time, especially in sectors where cash flows are directly aligned with timely consumer payments.

[1] KNIME AG. **KNIME**. Version 4.7.0, out.2022. <https://www.knime.com>. Zurich, Switzerland. Access October 2022

Water is a fundamental or essential asset, i.e. something that consumers cannot give up, even if their income falls or the price rises. When it chooses to do so, it will delay as long as possible, as slowly as possible (Ferguson, 1999; Varian, 1996). Delay is measured in days overdue (variable Delay). When it greater 50 days, Default condition (variable default=1).

Systems that learn from data and make decisions have become an increasingly concrete reality, with human supervision or totally independent, AutoML (He, 2021). The effort for complete independence is great, but the benefit is worthwhile. The effort is unique, but the benefit is continuous over time (Prasad, 2021).

Real phenomena, including economic phenomena, can be expressed using complex equations. Linear Regression techniques reduce them into simpler systems but maintain the relationships of equality and basic properties (Hill & Griffiths, 2011). The road to default is one of these phenomena. Various algorithms use these properties. They divide this data using regression rules into successively smaller sets, like a tree, called regression trees, obtaining satisfactory forecasting results, such as Regression Tree, XGBoost Tree and H20 (Blanquero et al., 2022).

There are several metrics for validating the models produced by algorithms based on regression, the main ones being: MAE, MSE, Kappa´s Coefficient and R^2 (coefficient of determination) varies between 0, which explains nothing, and 1, which would explain everything (Hair et al., 2009).

The growing importance of data-driven solutions and the prediction of defaults have driven the adoption of machine learning techniques in financial analysis. Kumar and Ravi (2021) emphasize how methods such as neural networks and support vector machines offer superior accuracy compared to traditional methods in predicting default. This not only enables more effective resource allocation for risk mitigation but also optimizes collection strategies by identifying potential defaulters before issues escalate. Thus, companies can tailor their approaches and reduce the negative impact of default on their financial health.

Data was extracted from J-TECH's SANSYS, specific version for CAJ[2], relating to charges (invoices) due between June/2020 and May/2022. The reference (referencia variable) of competence is from May/2020 to April/2022. The data extraction date is **16 December 2022.** In fact, there were 3.6 million payment records in this period, only 154,000 were in default. The Defauters (more of 50 days of Delay) represents 4.21% of total of records.

The procedure pivoted around reference by user ID.Analysing the Delay variation, the variables that proved relevant to the forecasting system were: *referencia, delay and default*.

RESULTS AND CONCLUSIONS

To say that water and hygiene are essential is almost repetitive, but the need for investment is still very high (Brazil, 2020). Defaults compete for resources with investments, and in some cases with costs themselves. On the other hand, current actions are reactive, stressful and potentializing conflicts. Preventive action is desirable, but the first step is to identify who will or will not become defaulters. In order to take preventive action, it is necessary to predict who will go into Default.

AI using machine learning and algorithms based on linear regression is capable of dealing with the issue. In fact, there are several models derived from the extracted data, the best model being based on the Regression Tree algorithm (Table 1). Machine learning is fundamental to this process, as these

[2] https://aguasdejoinville.com.br/sansys

prediction models need to be able to self-adjust, as their application will modify the reality on which they act.

By arranging the data referenced by the expiry date in this way and in this order, the analysis was performed using an algorithm based on linear regression. The result was a dichotomous variable, determining whether the consumer would go into Default or not.

The algorithm tests were carried out using AutoML tools with Knime software, with the main result criterion being the coefficient of determination. 80% of the records were used for training and 20% for running and testing the model. The algorithm with the best results was Regression Tree, with a R² of 0.9835. The tests with the reserved data (20%) for the generated model obtained excellent performance, generating a final table that informs each enrolment whether that consumer will go into Default or not.

By combining AutoML and linear regression algorithms with the extracted data, an analysis model was obtained that makes it possible to determine with 98.35 per cent assertiveness whether a given user or group of users will enter the Default condition, using only payment from the last 24 months (table 1).

Table 1 - Results AutoML (Knime)

Métricas ---->	R²	MAE	MSE	RMSE
Regression Tree	0.983470281619934	0.003857204239125	0.003203961643707	0.056603547978082
XGBoost Tree Ensemble	0.969001895845820	0.032223872277795	0.006008374399010	0.077513704588348
H20 Auto ML	0.960341105630161	0.040099903575008	0.008010881993114	0.089006348177873
Gradient Boosted Trees	0.841638819599762	0.101689320067466	0.030695208241798	0.175200480141459
Random Forest	0.744456912142841	0.151297677546798	0.049532014580241	0.222557890402118
Linear Regression	0.252559558326496	0.303324995456660	0.144876667043840	0.380626676736984
H2O Generalized Linear Model	0.252449554911941	0.304117240052200	0.144897989047810	0.380654684783742
XGBoost Linear Ensemble	0.082110589846192	0.369729791549453	0.177914855972182	0.421799544663404
Polynomial Regression	-0.000000001287769	N/A	N/A	N/A

Source: Authors

In this way, preventive measures can be adopted that avoid conflicts for both parties. This is before the situation of irremediable conflict occurs, the suspension of water supply services. Eliminating default completely is an intangible goal, but the problem must be minimized as much as possible. According to Wiener (1950), providing a more human use of data and technology.

Furthermore, the integration of advanced machine learning techniques within this analytical framework offers the potential for continuous improvement in predictive accuracy over time. By utilizing feedback loops and retraining the model with new data, the system can adapt to changing consumer behaviours and external economic factors, enhancing its predictive capabilities. This dynamic approach not only supports real-time decision-making but also aligns with strategic goals for risk management and resource optimization. As Jansen and Ritschl (2022) suggest, iterative learning processes in predictive models can lead to a significant reduction in misclassification rates, thereby bolstering confidence in pre-emptive action plans and long-term financial sustainability.

ACKNOWLEDGEMENTS

Special thanks to our mentor, Alexandre Leopoldo Gonçalves, for encouraging the pursuit of goals; and CAJ for providing the data for this study.

BIBLIOGRAPHIC REFERENCES

Blanquero, R. et al. (June 2022). On sparse optimal regression trees. *European Journal of Operational Research*, *299*(3),1045-1054. Retrieved July, 28, 2023, from https://doi.org/10.1016/j.ejor.2021.12.022

Borges, M.C.P., et al. (2022). The Brazilian National System for Water and Sanitation Data (SNIS): Providing information on a municipal level on water and sanitation services. *Journal of Urban Management*, 11, 4, 530-542

Brazil (2020, July). Law Nº 14,026, of 15 July 2020. Federal Government of Brazil (Presidency of the Republic, General Secretariat, Sub-Cabinet for Legal Affairs), Retrieved December 11, 2022 from http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/L14026.htm

Cui, W., & Kaas, L. (2021). Default cycles. *Journal of Monetary Economics*, *117*, 377-394. https://doi.org/10.1016/j.jmoneco.2020.02.001

Cui, W. (2022). Macroeconomic Effects of Delayed Capital Liquidation. *Journal of European Economic Association*, *20*(4), 1683-1742. https://doi.org/10.1093/jeea/jvac023

Ferguson, C.E. (1999). Microeconomia. Brazil, Rio de Janeiro: Forense Universitária

Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Multivariate Data Analysis (6th ed., p. 688). São Paulo: Bookman

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 1-27. https://doi.org/10.1016/j.knosys.2020.106622

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2011). Principles of econometrics (4th ed., p. 784). New Jersey, USA: *John & Sons*

Jansen, F., & Ritschl, A. *(2022).* Iterative Learning in Predictive Models: Enhancements in Accuracy. Computational Economics Journal.

Kiss, F., Schlichting, M., & Brochet, S. (2018). Credit Risk Analysis in Consumer Markets. Journal of Banking & Finance.

Kumar, V., & Ravi, V. (2021). Machine Learning Approaches for Banking Applications. Information Systems Frontiers.

Prasad, V.V. et al. (2021). Automating water quality analysis using ML and auto ML Techniques. Retrieved June 18, 2023, from http:// doi.org/10.1016/j.envres.2021.111720

Shaik, A. R., Boddu, D., & Samuel, P. (2020). Economic Impacts of Consumer Default. International Journal of Financial Studies.

Sinharoy, S. S., Pittluck, R., & Clasen, T. (2019). Review of drivers and barriers of water and sanitation policies for urban informal settlements in low-income and middle-income countries. *Utilities Policy Journal*, *60*, 2-8. https://doi.org/10.1016/j.jup.2019.100957

Varian, H.R. (1996). Intermediate Microeconomics**:** A Modern approach (4th ed., p. 650). USA: W.W. *Norton & Company Incorporation*

Wiener, Norbert. (1950). The human use of human being; USA: Houghton Mifflin Company.