PAYMENT DEFAULT PREDICTION: A potential and practical use of Artificial Intelligence in the public sector.

Tomelin,H.H.S.. htomelin@outlook.com, Oliveira Junior, S.M.. sidney.marques.mp@gmail.com, Zschoronack,T.. thiago.zschornack@gmail.com, Gonçalves,A.L.. a.l.goncalves@ufsc.br, Maia,R.G.B.. rafamp@gmail.com

Highlights:

- Reduction in losses due to defaults.
- conscious action versus instinctive reaction.
- A unique variable, payment variation, can predict defaulter in time series data set.

Keywords: Prediction; Sanitation; Default

INTRODUCTION

Basic sanitation is essential for human development. Its absence lowers the level of health of the population. It can be seen as a set of actions and services that provide adequate, organised and equitable hygiene and drinking water (Capone, 2020). In Brazil, services are priced individually (Borges et al., 2022). These tariffs provide the majority of funding. (Brasil, 2022).

Defaults erode these revenues, which are essential for funding, and overburden defaulters (Cui, 2022). In Jaraguá do Sul, a medium-sized city with good hygiene and HDI indices, the loss from bad debts is around 4 percent of sales. By way of comparison: This amount over 4 years is equivalent to installing a sewage collection system for 14,000 people.

The action that is usually taken in relation to non-payers: notification and disconnection of services (after 50 days of delay), known as the "default condition". This reactive strategy is time-consuming, ineffective, and places a double burden on defaulters. In order to adopt a new preventive strategy, it is necessary to identify the non-payer before he defaults.

The combination of mathematical and artificial intelligence for time series analysis can provide a solution to the first and essential step in solving this question.

METODOLOGHY

According to Varian (1996), default is defined as when the consumer fails to pay on time (disengagement) or even fails to pay at all, a cost which is considered systemic (independent of the organization's internal management). Cui (2022) reinforces this concept, highlighting the financial and institutional stress that arises in organizations.

Water is a fundamental or essential asset, i.e. something that consumers cannot give up even if their income falls or the price rises. If they decide to do so, they will delay as long and as slowly as possible (Ferguson, 1999; Varian, 1996). In this article, default is measured in days overdue, which is referred to as delay (also a variable Delay).

According to Russell and Norvig (2003), artificial intelligence (AI) can be analysed along several dimensions, including thought processes, reasoning and behaviour. AI can be understood as "the art of creating machines that perform functions that require intelligence when performed by humans" (Kurzweil, 1990). In this way, AI gives machines capabilities that are inherent to humans, such as visual perception and analysis, natural language processing, pattern recognition, decision making, and many other examples.

Systems that learn from data and make decisions are becoming an increasingly tangible reality, with human oversight or completely independent, automated machine learning or AutoML (Russel & Norvig, 2003). It is therefore distinctive by techniques that allow computers to automate the construction of analytical models (based on mathematical models) from any data type (He, 2021).

Several algorithms use linear regression as a mathematical model. According to Hill&Griffitths (2011), it consists of constructing an interpretation of a complex function into a simple function of the classical linear type. The algorithms divide this data using regression rules into successively smaller sets, like a tree, which is why they are called regression trees, obtaining satisfactory forecasting results, such as LGBM and Random Forest. It is worth mentioning that the software library chosen for this work is PyCaret4, which, according to Ali (2020), is an open-source Python® library aimed at automating ML tasks.

There are several metrics for validating the models generated by regression-based algorithms, the main ones being: MAE, MSE, RMSE, MAPE, and **R²** (coefficient of determination) which is the most significant (Hair et al., 2009) (Padhma. 2021).

Data was extracted from J-TECH's SANSYS[1], specific report FAT0022 (J-TECH, 2022), relating to charges (invoices) due between from May/2020 and May/2022. The extraction took place on 21 October 2022. The difference between the payment date and the due date in days was recorded in *Delay* (records not yet discharged were taken as the date of extraction): Early, negative; Punctual, zero; Late, positive. The database has 1,142,029 records.

RESULTS AND CONCLUSION

A first analysis of the data reveals a scenario of 10.43% of records in default status, with a delay of 50 days or more, as shown in Table 1.

Table 1 - Delay of Payments records

Days	Occurrences	% total
<= 0	702,175	61.48
>= 1	190,200	16.65
>=15	80,374	7.04
>=31	50,200	4.40
>=50	119,080	10.43
Total	**1,142,029**	**100.00**

Source: Authors

The variables that proved to be relevant to the forecasting system, focusing on the analysis of delay variation, were: *referencia* (measurement reference) from 1 (May/20) to 25 (May/22); *delay* (integer) and *default* (default condition: on=1, off=0) for each reference. After pivoting around the *referencia* aggregated by *matricula*, the database was reduced to 47,441 records. By arranging the sequence, referenced by *referencia* (1 to 25) in this way and in this order, the analysis was performed using an algorithm based on linear regression. The result determined whether the consumer would default or not.

After analysing the data, it is observed that 10.43% of the records are in default status, with delays of 50 days or more. This scenario was crucial in guiding the choice of prediction algorithms, as identifying

[1] *SANSYS*. Retrieved April 21, 2023, from www.sansys.com.br

relevant variables such as the time reference and the default status was essential. By aligning the information around the reference, aggregated by ID, the dataset was reduced to 47,441 records, enabling a more robust analysis through linear regression. This simplified and optimized methodology contributed to developing an efficient model capable of predicting default patterns with remarkable accuracy.

To determine the best algorithm for predicting defaulters, AutoML was run, the data was split in a balanced way, with 80 per cent (37,952 records) for training and 20 per cent (9,489 records) for testing the model. The best result for R^2 was 97.09% for Light Gradient Boosting Machine (LGBM) as can be seen from Table 2.

Table 2 - Algorithm test results with AutoML by Pycaret®

Abrev.	Model Name	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT(sec)
lightgbm	Light Gradient Boosting Machine	0.0315	0.0069	0.0829	0.9709	0.0581	0.0516	0.7710
rf	Random Forest Regressor	0.0207	0.0070	0.0837	0.9703	0.0618	0.0260	37.4510
et	Extra Trees Regressor	0.0405	0.0113	0.1064	0.9522	0.0774	0.0569	12.9140
dt	Decision Tree Regressor	0.0149	0.0149	0.1218	0.9372	0.0844	0.0174	0.6290
knn	K Neighbours Regressor	0.0696	0.0405	0.2011	0.8293	0.1311	0.1531	2.8800
gbr	Gradient Boosting Regressor	0.1725	0.0589	0.2426	0.7516	0.1605	0.2626	7.9620

Source: Authors

Running the same algorithm yielded a result of 97.15 per cent assertiveness, which is consistent with the training and testing stages. This is show in Table 3.

Table 3 - Result LGBM-based algorithm execution

Sigla	MODELO	MAE	MSE	RMSE	R^2	RMSLE	MAPE
lightgbm	Light Gradient Boosting Machine	0.0317	0.0069	0.0828	0.9715	0.0578	0.0512

Source: Authors

The tests conducted with AutoML and the subsequent implementation of the Light Gradient Boosting Machine (LGBM) resulted in a determination coefficient (R^2) of 97.09%, highlighting the method's effectiveness in predicting default. The execution of the same algorithm confirmed an accuracy of 97.15%, consistent with the training and testing phases. These metrics underscore the reliability of LGBM, which outperformed other models tested, such as Random Forest and Extra Trees Regressor. The final structuring of the model allows for efficient integration of results into a simple output table, which classifies users as defaulters or non-defaulters, facilitating its use in credit management systems.

A single explanatory variable, through its variation, revealing the behavioural inflection (Varian, 1996), combined with computer algorithms based on linear regression and AutoML are able to generate satisfactory models for predicting which users will default. The final processing result based on the model built using the LGBM algorithm should have an output table with two columns: "matricula" (ID of the user), and "default", which indicates whether the user will default on the next reference, where 1 indicates (default) and 0 (no default). A simple data format can be efficiently integrated.

The logical sequence of this research: Development of an API, available on the internet, which can use a lower time series and a return with probability of the user going into default.

If the prediction is made before the situation of irretrievable conflict arises, the suspension of water supply services and its consequences can be avoided. For both sides, the consequences are terrible. A complete elimination of the problem is not realistic, but a significant reduction is.

ACKNOWLEDGMENTS

Special thanks to our mentor, Alexandre Leopoldo Gonçalves, for encouraging the pursuit of goals.

Thanks for Jaraguá do Sul´s SAMAE for providing the data for this study.

Thanks for Santa Catarina University Scholarship Program (Uniedu) of the Secretary of Education.

REFERENCES

Ali, M. (2020, April). An open source, low-code machine learning library in Python. Pycaret Docs. Retrieved April 2020 from https://pycaret.gitbook.io/docs

Brasil (2022, December). Governo Federal do Brasil (Secretaria Nacional de Saneamento Ambiental - SNASA- do Ministério do Desenvolvimento Regional), Retrieved December 11, 2022, from http://app4.mdr.gov.br/serieHistorica

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.

Capone, D., Cumming, O., & Nichols, D. (2017). Water and sanitation in urban América. *American Journal Public Health Association*, *110*(10), 1445-1578. Retrieved October 2020, from https://ajph.aphapublications.org/doi/epdf/10.2105/AJPH.2020.305833

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

Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise Multivariada de Dados* (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

Kurzweil, R. (1990). *The Age of Intelligent Machines* (p. 580). MIT Press

Padhma, M. (2021). *End-to-End Introduction to Evaluating Regression Models*. Retrieved April 17, 2023, from https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd Ed., P. 1132). New Jersey: Prentice Hall

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