Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n1p225

A Two-Stage Model for Analyzing Customer Service Costs

By Barzizza; Disegna; Fanesi

15 March 2025

Electronic Journal of Applied Statistical Analysis
Vol. 00, Issue 00, August 2013, 225-237
DOI: 10.1285/i20705948v18n1p225

A Two-Stage Model for Analyzing Customer Service Costs

Elena Barzizza [*a], Marta Disegna[b], and Alessandro Fanesi[b]

[a] *University of Padova, Department of General Psychology, Via Venezia, 8, 35131, Padova*
[b] *University of Padova, Department of Management Engineering, Stradella San Nicola, 3, 36100, Vicenza*

15 March 2025

In industrial contexts where managing customer service costs is critical, accurately predicting and analyzing these costs presents a significant challenge, particularly when dealing with zero-inflated count data. This study proposes a Two-Stage Machine Learning approach that extends the traditional hurdle model, offering enhanced flexibility and adaptability to complex data structures without compromising interpretability. Through a real-world case study in the cleaning service sector focused on one-time service purchases, the proposed method identifies key cost drivers and provides actionable insights into customer behavior. This research advances the field by presenting a highly effective method for analyzing zero-inflated data, outperforming popular models based on Poisson distribution. Simultaneously, it addresses practical business needs by supporting data-driven strategies to optimize operational resources and manage customer costs more effectively.

keywords: Machine Learning, Two-Stage, Zero-Inflation, Business Case.

1 Introduction

In real industrial contexts where customer service costs are a significant concern, accurately predicting and analyzing these costs is crucial. This study focuses on addressing the challenge of understanding and predicting customer costs, with particular emphasis on identifying the key factors influencing these predictions. The specific case study explored focuses on a customer cleaning service company that primarily deals with customers purchasing one-time services, where payment is made upfront rather than through

[*]Corresponding authors: elena.barzizza@unipd.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

recurring subscriptions (e.g., monthly or annual fees), though customers may hold both service types. A notable challenge in this context arises from the presence of zero counts, which occur when a portion of the population exhibits no measurable activity related to the cost metric under investigation. Addressing this issue requires an appropriate methodological approach.

When the frequency of zero counts exceeds what is expected under a Poisson model, the phenomenon of interest is said to exhibit zero inflation. Indeed, distributions such as the Poisson are unable to adequately model the remaining portion of counts different from zero, leading to the problem of overdispersion when this portion of the counts displays significant variability. For this reason, the literature offers a wide range of methodologies developed with the aim of modeling this unobserved heterogeneity that the Poisson model fails to capture.

Following Cameron and Trivedi (2013), the models used to address the problem of zero-inflated count data can be grouped into four different families: mixture models for unobserved heterogeneity; waiting time distribution models; flexible methods allowing for the modeling of both overdispersion and/or underdispersion; finite mixture models.

Recently, the literature has focused on the hurdle model, the zero-inflated model, and their related extensions.

The hurdle model (Mullahy, 1986) can be interpreted as a two-stage model based on the assumption of independence between the distribution generating zeros and the distribution generating the positive values. Thus, it is a two-component mixture model consisting of a point mass at zero followed by a zero-truncated count distribution for the positive observations. For the count part of the hurdle model a common choice is to use a negative binomial distribution (Shankar et al., 2022; Oyhenart, 2020; Bhaskar et al., 2023).

The zero-inflated model is a mixture of two components: a degenerate distribution at zero and a untruncated count distribution, so the assumption of independence is not applied. As a result, zero counts are interpreted differently. For the count part of the zero-inflated model a common choice is to use the Poisson distribution (Liu et al., 2021; Bracamontes et al., 2020; Chaves and Friberg, 2021) or the Negative Binomial distribution (Rao and Babu, 2021; Chaves and Friberg, 2021; Bhaskar et al., 2023).

Both hurdle and zero-inflated model therefore rely on the assumption that the population can be divided into two groups: those who are never at risk of experiencing the event and those who may show a positive count.

In the context of Machine Learning, Abraham and Tan (2009) emphasize in their work how the abundance of zeros affects the adaptation of Machine Learning models, making them unable to correctly predict both zeros and positive values. Another issue is related to the distortion caused by zeros in the calculation of error metrics, making it difficult to estimate the predictive capacity of models Rozanec et al. (2023). For this reason, the literature includes many methods developed to address this problem. For example, Abraham and Tan (2009) tackled zero inflation by extending the logic of zero-inflated models to a semi-supervised Machine Learning context. Their approach follows a two-stage framework: first, a classification model determines whether a value is nonzero; then, a regression model estimates its magnitude, treating both stages as independent.

Following the works of Hu et al. (2022); Rozanec et al. (2022, 2023); Krasniqi et al. (2023); Xu et al. (2024), the main advantages of considering two independent stages in Machine Learning are:

1. The ability to optimize the two problems separately using specific error metrics. This allows for identifying where the models perform well or poorly and making improvements as needed.

2. The ease of describing the data generation process. Through model interpretation, it becomes possible to identify which factors have a greater impact on the classification problem versus the regression problem.

The approach most commonly pursued is to consider classification and regression models independently and for each task selecting the one with the best performance (Hu et al., 2022; Rozanec et al., 2023, 2022; Xu et al., 2024).

1.1 Scope of the work

This work focuses on understanding customer behavior by modeling individual purchase processes as a two-stage decision framework. In this context, the decision-making process consists of two sequential stages: first, the classification problem about the decision to purchase a service, and second, the regression problem about the decision regarding the amount to spend on it Pudney (1989). Such a structure naturally aligns with two-stage models, such as the hurdle model, which explicitly accounts for this two-step process. However, the hurdle model must satisfy certain assumptions, such as the distribution of the two components and the structure of the function that relates the response variable to the covariates. While effective in many settings, these assumptions may limit its flexibility when applied to complex, real-world data. To address this limitation, we adopt a Machine Learning framework within a two-stage context to analyze zero-inflated count data, offering greater adaptability to the data structure and reducing reliance on strong distributional assumptions. This method can be viewed as an extension of the hurdle model within the Machine Learning context, as the two stages operate independently.

Through a real-world case study in the customer cleaning service sector, this study may offer both theoretical and managerial insights. The findings shed light on the key factors influencing customer costs and their predictions that provides actionable knowledge for businesses. From a practical perspective, this analysis supports data-driven decision-making by identifying cost-driving patterns, thereby enhancing customer service operations. Therefore, this study highlights how incorporating two independent stages in Machine Learning can improve the analysis of zero-inflated data in customer service applications, offering practical improvements. The proposed Two-Stage Machine Learning method is compared against most popular Poisson models in terms of predictive error as they are often used as a benchmark technique for models validation Xu et al. (2024).

The paper is structured as follows: Section 2 focuses on the methodological development of the Two-Stage Machine Learning approach, while Section 3 illustrates its application using real business case data. Final remarks are discussed in Section 4.

2 Methodology

Considering y as the zero-inflated count target variable to be modeled, the two stages are considered independently. The assumption of independence between the two stages enables the separation of the analyses using distinct models for classification and regression.

Initially, a train-test split is performed. Several classification models are trained to predict a binary target variable, which takes the value 0 if $y = 0$ and 1 if $y > 0$. The performance of these classification models is evaluated and compared using the test set, allowing the selection of the best classification model. Similarly, several regression models are trained on the subset of the training data where $y > 0$ to predict the count y. The test set is then used to identify the best-performing regression model. The selected classification and regression models are then combined to make predictions for new observations. Specifically, the classification model first determines whether y is greater than 0. If $y = 0$, the predicted count is set to 0. Otherwise, the regression model is used to predict the exact count value. We used therefore this approach to evaluate the overall prediction error on the test set.

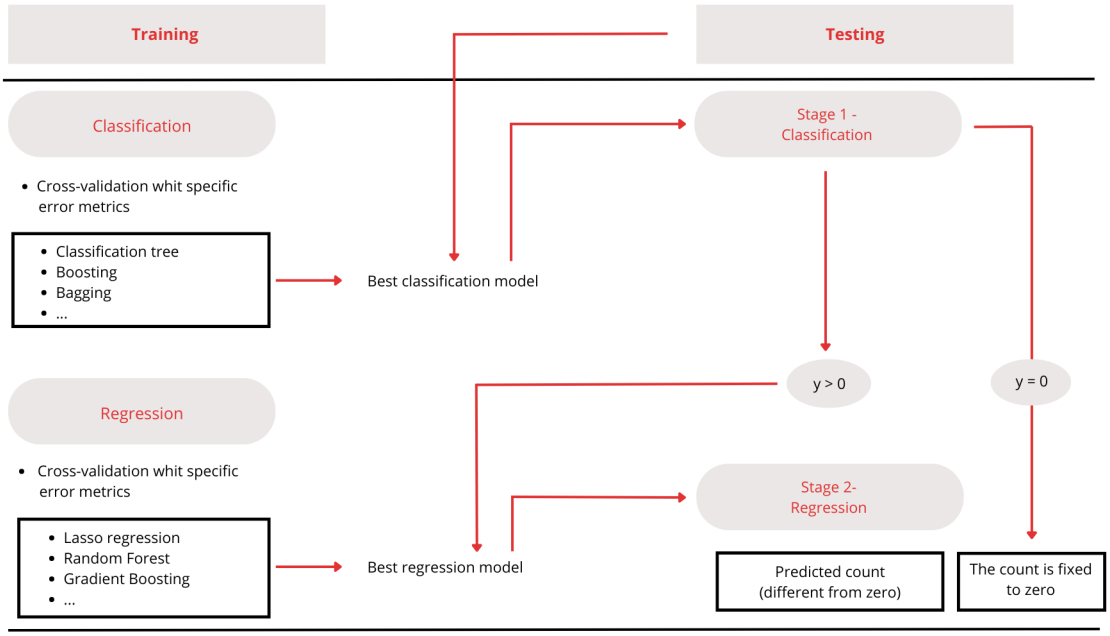

Figure 1: The Two-Stage Machine Learning Approach. The variable y represents the zero-inflated count target variable.

Figure 1 presents the two-stage modeling logic described above. Finally, the best models can be interpreted through the SHAP value metric (Strumbelj and Kononenko, 2014).

The error metrics used to select the best classification model are: misclassification error rate, false negative rate, false positive rate, F1 score and sensibility (Xu et al., 2024). While to identify the best regression model: the MSE on the logarithmic scale; the SQRTMSE and MAE calculated on the original scale of the data, truncated by removing 0.05% of values from both the right and left tails of the prediction residuals. This approach provides a robust error metric against outliers in y, reflecting the model's performance in predicting the average behavior of the response. Additionally, the MAE on the original scale without truncation is reported to account for outliers and thus reflect the entire population.

3 Case study

This case study examines customer behavior and service utilization patterns within a cleaning service company over a 26-year period (1997–2023). The statistical units are the customers who signed a contract with a cleaning service company between 1997 and 2023. These customers are those who purchase one-time service, where payment is made only once, as opposed to subscription-based service that require periodic payments such as monthly or annual fees. In any case, it is possible that a customer with one-time service is also a customer with subscription-based services, but this is not necessary.

For measuring the cost of each customer, the amount of time operators must dedicate to customer requests is used. This is measured by the weighted average of the number of calls and opened tickets generated by the customers, denoted as number of contacts (y).

The primary objectives of this analysis are twofold: first, to accurately predict the number of contacts, and second, to identify the key factors influencing these predictions. These goals are achieved using the Two-Stage Machine Learning approach detailed in Section 2.

The factors considered are 14 and relate to both customer characteristics and contract details. Table 1 provides the complete list along with a brief description of each factor. The `R` software is used to develop all the analysis and graphs.

In Figure 2, the histogram on the left shows the distribution of the target variable, while the histogram on the right focuses on values of y less than 100. Figure 3 presents boxplots of the target variable in the same order, considering all its values and only those where y is less than 100. The vast majority of observed number of contacts are close to 0.

From the analysis of both histograms and boxplots, it emerges that the distribution of the target variable y (i.e., number of contacts) for customers with one-time purchases is highly skewed, with a high percentage of values close to zero and a few positive outliers. Specifically, the percentage of customers with zero contacts, meaning those who have never opened a ticket or made a call, is approximately 13%. Obviously, the marginal distribution of y is being evaluated, which could differ significantly from the conditional distribution with respect to the covariates. In any case, it is reasonable to assume that, given the nature of the response, there will also be a segment of customers who have

Table 1: List and description of factors considered in the case study.

Factors	**Description**
ServiceArea	Geographical area of the service
ClientType	1=individual; 0=company
NumServices	Number of active service agreements
Status	Status of the agreement (active, inactive, or mixed)
Activation	Activation period of services (before 2020, before 2023, between 2020 and 2023)
NumCleanings	Total number of cleaning services performed for the client
ServiceTypes	Number of different types of cleaning services used
ServiceCombo	Most frequent combinations of cleaning services
AveRevenue	Average revenue generated by the client
Duration	Duration (in years) from the first to the most recent cleaning service
AssignedRep	Assigned commercial representative (or partner) for the client
Partner	Presence of a commercial partner during service agreement discussions
Payment	Method of payment used by the client
Rent	Indicates whether the customer has subscribed to a subscription-based service

not contacted the company and another segment who have contacted it multiple times, even conditionally. This implies that, in general, one can expect to observe a substantial proportion of zeros as well as a wide range of count values.

The Two-Stage Machine Learning approach is implemented in this analysis. Specifically, 70% of the data (7133 observations) is allocated for training the model, utilizing a five-fold cross-validation procedure to identify the best performing machine learning model. The remaining 30% of the data (3058 observations) is used to compare the performance of the Two-Stage Machine Learning approach with its competitors.

3.1 First stage: classification

In the classification stage, the target variable 'number of contacts' is transformed into a binary variable that takes the value 0 if the number of contacts is 0, and 1 if the number of contacts is greater than 0.

We decide to apply a range of models from simple to more complex among those most commonly employed in classification tasks. This approach allows us to compare different models and select the most appropriate solution for the problem at hand. Our selection includes: Lasso Regression both with only marginal effects (Lasso) and with interaction effects (Lasso-int), Classification Tree (Tree), Boosting applied both in the version with Classification Trees of depth one (Boost - Stumps) and with Classification Trees grown to a depth of two (Boost), Bagging, Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), Gradient Boosting (GBM), and Support Vector Machines (Svm). Each

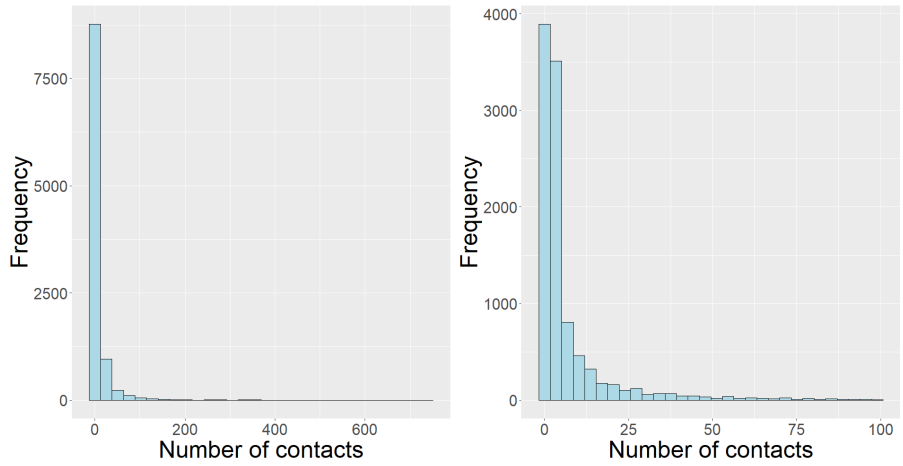

Figure 2: Marginal distribution of the target variable y (number of contacts for customers with one-time purchase services).

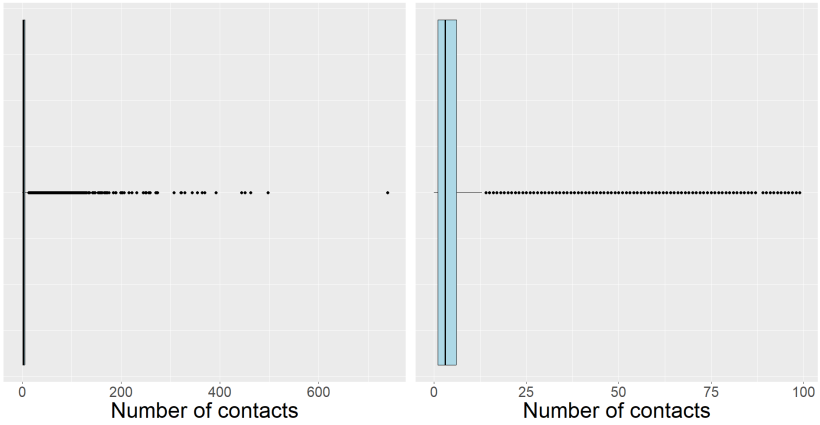

Figure 3: Boxplot of the target variable y (number of contacts for customers with one-time purchase services).

model offers distinct advantages in handling different data characteristics such as non-linearity, high dimensionality, and interaction effects.

In Table 2, the values of the performance indicators and the corresponding standard errors, calculated using the cross-validation procedure, are reported for the various implemented classification algorithms.

In this case, the best model is the RF, as it maintains a low false positive rate (28%) and has a false negative rate of approximately 16%. Additionally, the F1 score is 88%, indicating that the model predicts positives well.

Table 2: Error metrics (and corresponding standard errors) calculated for the machine learning models used in the classification stage.

Model	**MEr**	**FPr**	**FNr**	**F1**
Lasso	0.1704 (0.0370)	0.1704 (0.0370)	0.2449 (0.0201)	0.8398 (0.0096)
Tree	0.2185 (0.0729)	0.2185 (0.0729)	0.2634 (0.0687)	0.8204 (0.0370)
Lasso - int	0.2187 (0.0130)	0.3960 (0.1621)	0.1737 (0.0418)	0.8579 (0.0098)
Boost	0.1818 (0.0059)	0.2390 (0.0737)	0.1668 (0.0135)	0.8800 (0.0025)
Boost - Stumps	0.1818 (0.0059)	0.2390 (0.0737)	0.1668 (0.0135)	0.8800 (0.0025)
Bagging	0.2012 (0.0127)	0.2097 (0.0345)	0.1985 (0.0183)	0.8644 (0.0074)
RF	**0.1829** (0.0048)	**0.2784** (0.0464)	**0.1585** (0.0175)	**0.8803** (0.0043)
MARS	0.2002 (0.0077)	0.2490 (0.0461)	0.1874 (0.0170)	0.8666 (0.0050)
GBM	0.1990 (0.0078)	0.2724 (0.0530)	0.1798 (0.0186)	0.8683 (0.0056)
Svm	0.1872 (0.0045)	0.2945 (0.0416)	0.1599 (0.0136)	0.8777 (0.0034)

3.2 Second stage: regression

In the second stage, multiple regression models are trained on the portion of the training set where the 'number of contacts' is greater than zero. The following Machine Learning models are used: Ridge Regression (Ridge - min, Ridge - 1se), Lasso Regression (Lasso - min, Lasso -1se, Lasso - min - int), Regression Tree (Tree - min, Tree -1se), Multivariate Adaptive Regression Splines (MARS), Generalized Additive Model (GAM), Random Forest (RF), Neural Network (NN), and Gradient Boosting GBoost. In particular, Ridge, Lasso, and Regression Tree models are considered both in the case where their respective smoothing parameters correspond to the minimum value (min) and according to the one standard error (1se) rule to consider more parsimonious models that may better generalize the phenomenon being analyzed.

We employ a diverse range of regression models from simpler regularized linear approaches to more complex algorithmic techniques. This selection represents models with varying degrees of flexibility and interpretability, allowing us to effectively capture different aspects of the underlying data patterns. This comprehensive approach enables us to identify the most effective model for predicting the target variable y.

Table 3 reports the values of the various performance indicators and their associated standard errors for the adopted models.

In this case, the best model is the RF, as it maintains a low MSE similar to that of Gradient Boosting, while also exhibiting low truncated SQRTMSE, MAE, and truncated MAE.

Table 3: Error metrics (and associated standard errors) calculated for the adopted regression models.

Model	**MSE**	**SQRTMSE-truncated**	**MAE-truncated**	**MAE**
Ridge - min	0.5939 (0.0198)	4.9424 (0.1158)	3.0422 (0.0740)	6.5244 (1.2327)
Ridge - 1se	0.6015 (0.0177)	4.9858 (0.1343)	3.0413 (0.0680)	6.4210 (1.0158)
Lasso - min	0.5937 (0.0198)	4.9502 (0.1208)	3.0416 (0.0798)	6.4602 (1.1672)
Lasso - 1se	0.6022 (0.0184)	5.0225 (0.1201)	3.0619 (0.0754)	6.2401 (0.9307)
Tree - min	0.6246 (0.0226)	5.4039 (0.1552)	3.2244 (0.0697)	5.9571 (0.5464)
Tree - 1se	0.6361 (0.0247)	5.4611 (0.1792)	3.2451 (0.0750)	5.9867 (0.5318)
Lasso - min - int	0.6720 (0.0266)	5.4054 (0.2518)	3.2216 (0.1284)	6.1226 (0.4426)
MARS	0.5935 (0.0223)	4.9356 (0.1746)	3.0057 (0.0850)	5.2363 (0.3372)
GAM	0.5662 (0.0246)	4.7846 (0.1208)	2.9409 (0.0866)	5.2959 (0.4085)
RF	**0.5313** (0.0162)	**4.6631** (0.0788)	**2.8734** (0.0533)	**5.2518** (0.4479)
NN	0.5599 (0.0252)	4.8804 (0.1728)	2.9754 (0.0966)	5.2511 (0.3873)
GBoost	0.5311 (0.0224)	4.8310 (0.1910)	2.9433 (0.1039)	5.3011 (0.5031)

3.3 Combination of models

The predictions from the two top-performing models in the classification and regression stages are then integrated. Conditional on the classifier's prediction that the customer will call or open a ticket, the number of customer contacts is predicted using the regression model.

3.4 Comparison of the Two-Stage Machine Learning model and Poisson models

Table 4 contains the error metrics calculated on the testing data for both the Two-Stage Machine Learning model and Poisson models. It shows that the Two-Stage Machine Learning model, composed by the combination of two Random Forest (RF), has all error metrics lower then all ZIP models. This results marks the increased performance reached by the Two-Stage Machine Learning approach over the Poisson models. It's important to notice that the ZINB model returned a computational error during the estimation phase, therefore it's not considered in the comparison.

The Two-Stage Machine Learning model proposed in this study demonstrates clear advantages over traditional approaches. Moreover it is possible to underline the key advantage in terms of model interpretation linked to the Two-Stage method. In fact Poisson models do not allow to treat independently the classification and the regression problem, producing a vague interpretation of the variables' effect in generating zeros or positive values. Using a Two-Stage Machine Learning model, it is also possible to understand which are the most important factors through the SHAP value. Figure 4

Table 4: Comparisons of the error metrics.

Model	**MSE**	**SQRTMSE-truncated**	**MAE-truncated**	**MAE**
Two-stage RF	393.997	3.8217	2.5908	4.9427
ZIP	1086.996	4.8844	3.1947	6.0079
HP	1087.974	4.8877	3.1926	6.0063
HNB	268017.6	4.5139	2.9951	15.3912

and Figure 5 shows the variable importance graphs respectively for the best classification model and the best regression model selected in the training data. The importance is measured by the average of the absolute SHAP value calculated for each factor. From these graphs it is possible to see how factors' importance changed from the classification model to the regression one, allowing a clear understanding of which factors influences mostly the decision to contact the company and the number of contact once the customer has already decided to contact the company.

To determine whether the number of contacts is greater than 0, the most relevant factor appear to be Status, Activation, Duration, and NumCleanings. On the other hand, for the second stage model the most important predictors are Rent, ClientType, NumCleanings, and NumServices. By comparing the two variable importance plots, we can observe some differences. In particular, we can highlight that the number of active service agreements is important for the regression task but not for the classification task. From a managerial point of view, knowing the importance of the different factors allows the company to identify the ones that can increase the probability of having at least one contact. Additionally, understanding the most important factors according to the regression model helps us understand how to minimize the number of contacts.

It is worth mentioning that, while the Two-Stage Machine Learning model offers the aforementioned advantages, it does come with a higher computational effort compared to traditional models. However, the trade-off is justified by the improved interpretability and performance achieved.

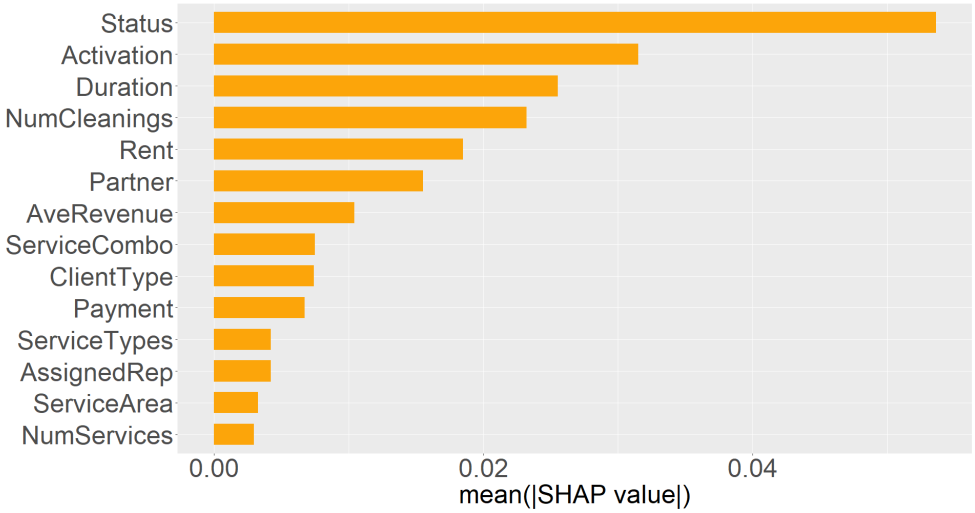

Figure 4: Variable importance graph of the best classification model (Random Forest).

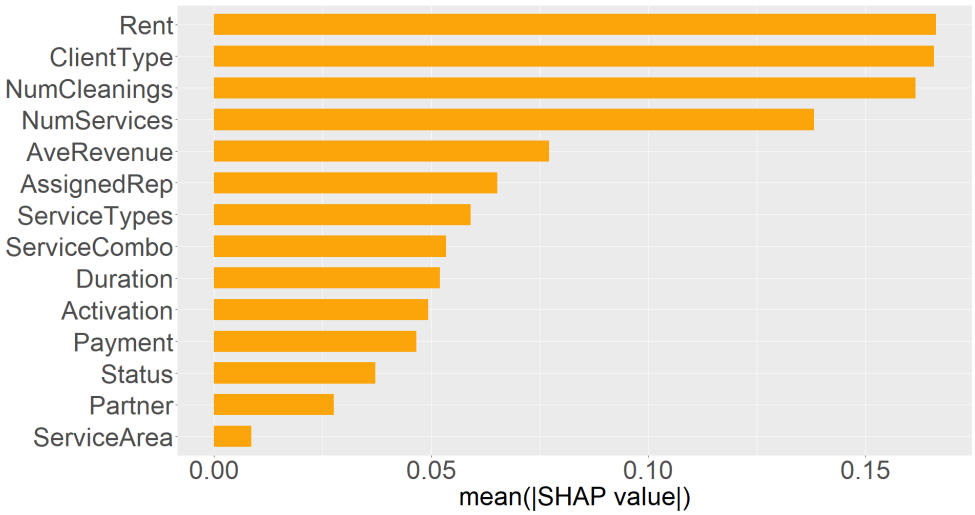

Figure 5: Variable importance graph of the best regression model (Random Forest).

4 Conclusion and further research

This paper addresses the critical challenge of predicting and managing customer service costs in real-world industrial contexts, focusing on zero-inflated count data, which introduces significant modeling complexities. From a theoretical standpoint, this study addresses the limitations of traditional models in handling zero-inflated count data. The paper introduces a Two-Stage Machine Learning approach, extending the hurdle model. This method treats the zero-count and non-zero count stages as independent, optimizing each stage separately and enhancing flexibility while maintaining interpretability. This

approach not only simplifies the process of interpreting results but also allows for a more detailed understanding of the factors influencing customer costs, divided into the two stages of the analysis, namely the classification and regression tasks.

From a practical perspective, the study offers actionable insights for businesses to predict and manage customer service costs more effectively. By identifying key cost drivers, companies can make informed decisions, optimize resources, and improve operational efficiency. Ultimately, this research promotes data-driven decision-making to enhance customer service operations and manage costs in a sustainable manner. In addition to contributing to zero-inflated count data prediction literature, this study reveals various paths for further research. As a next step of this research we will weaken the assumption of independence between the two steps without introducing selection bias.

References

Abraham, Z. and Tan, P.-N. (2009). A semi-supervised framework for simultaneous classification and regression of zero-inflated time series data with application to precipitation prediction. *2009 IEEE International Conference on Data Mining Workshops*. 644–649.

Bhaskar, A., Thennarasu, K., Philip, M., and Jaisoorya, T. (2023). Regression models for count data with excess zeros: A comparison using survey data. *Quantitative Methods for Psychology*, 19:1–13.

Bracamontes, C. G., Carrillo, T., Montealegre, J., Fradkin, L., Follen, M., and Mulla, Z. D. (2020). Analysis of count data in the setting of cervical cancer detection. *Journal of Investigative Medicine*. 68: 1196–1198.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

Chaves, L. F. and Friberg, M. D. (2021). Aedes albopictus and aedes flavopictus (diptera: Culicidae) pre-imaginal abundance patterns are associated with different environmental factors along an altitudinal gradient. *Current Research in Insect Science*. 1: 100.

Hu, X., Hu, J., and Hou, M. (2022). A two-step machine learning method for casualty prediction under emergencies. *Journal of Safety Science and Resilience*. 3: 243–251.

Krasniqi, D., Bardet, J.-M., and Rynkiewicz, J. (2023). Parametric and xgboost hurdle model for estimating accident frequency. *HAL open science*.

Liu, Z., Kemperman, A., and Timmermans, H. (2021). Correlates of frequency of outdoor activities of older adults: Empirical evidence from dalian, china. *Travel behaviour and society*, 22:108–116.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*. 33: 341–365.

Oyhenart, J. (2020). Zero inflated count regression for one year prediction of bovine trichomonosis in a compulsory control plan in la pampa, argentine. *Veterinary Parasitology: Regional Studies and Reports*, 20:100394.

Pudney, S. (1989). *Modelling individual choice: the econometrics of corners, kinks and holes*. Basil Blackwell.

Rao, G. B. and Babu, S. (2021). Roost-site selection and population assessment of gulls wintering along india's west coast reveals the importance of conserving coastal habitats. *Ornithological Science*, 20:161–174.

Rozanec, J. M., Fortuna, B., and Mladenic, D. (2022). Reframing demand forecasting: A two-fold approach for lumpy and intermittent demand. *Sustainability*, 14:92–95.

Rozanec, J. M., Petelin, G., Costa, J., Bertalanic, B., Cerar, G., Guvek, M., Papa, G., and Mladenic, D. (2023). Dealing with zero-inflated data: achieving sota with a two-fold machine learning approach. *arXiv.org*.

Shankar, S. V., Ajaykumar, R., Prabhakaran, P., Kumaraperumal, R., and Guna, M. (2022). Modelling of tea mosquito bug (helopeltis theivora) incidence on neem tree: A zero inflated count data analysis. *Journal of Agrometeorology*, 24:409–416.

Strumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665.

Xu, X., Ye, T., Gao, J., and Chu, D. (2024). Generalized hurdle count data models based on interpretable machine learning with an application to health care demand. *Computing*.