Journal of Kyzylorda Scholarly Review

P-ISSN:3006-0265 | E-ISSN:3007-4681. Published by the Open University of Kyzylorda
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY).
https://bulletin.ouk.kz/index.php/bulletin/index



MODELING ACADEMIC PRODUCTIVITY USING ZERO-INFLATION AND MACHINE LEARNING METHODS FOR COUNT DATA

Wakaa Ali Hadba
Independent research
wakaa2017@uokirkuk.edu.iq

Abstract. This research aims to investigate statistical models and machine learning techniques used to analyze numerical data suffering from overdispersion and a high proportion of excess zeros, common features in many fields such as insurance, healthcare, and the environment. In this study, we explore predictive modeling techniques for analyzing scientific productivity, measured by the number of articles published among biochemists. This study aims to evaluate and compare the predictive performance of classical zero-inflated regression models and modern machine learning algorithms in statistical data modeling, using academic productivity (number of published articles) as the main outcome variable. The dataset exhibits typical statistical data characteristics, including overdispersion and an excess of zeros. Therefore, several statistical models were considered, such as Poisson (POS), negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). As well, a collection of machine learning algorithms helps the use of the vector machines (SVMs), neural networks (NNs), random forests (RFs), and XGBoost were used to compare them. A number of evaluation metrics were used to assess the model performance:Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R 2, Akaike Information Criterion (AIC)Bayesian Information Criterion (BIC). Log Likelihood (LogLik), Dispersion Index.The results showed that while traditional models such as ZINB and NB effectively handled overdispersion and excess zeros, Random Forest outperformed all models in terms of prediction accuracy, achieving the lowest RMSE and MAE values. SVM and ZIP followed with average performance, while XGBoost and NN showed relatively higher error rates without tuning.The results indicate that while statistical models remain valuable for their interpretability and clear handling of data characteristics, tree-based machine learning models, such as Random Forest, offer superior prediction performance for count data when prioritizing model accuracy.

Key words: Zero-Inflation Models (ZIM), Machine Learning Algorithms, XGBoost, Random Forest, Neural Network.

Introduction

Highly dispersed numerical data are characterized by high variance relative to the arithmetic mean and a high proportion of zero values. This is a common statistical challenge in many applied fields, including insurance, medicine, social sciences, and the environment. The Poisson distribution is the most widely used distribution for counting data. Overdispersion occurs when the variance exceeds the mean, leading to model misspecification. Overdispersion and zero inflation are two common problems with the Poisson model, which assumes equal mean and variance of the number of events, and are often insufficient for such data. This shortcoming can lead to underestimated standard errors and misleading inferences[1]. These properties lead to poor performance of classical models, such as the Poisson model, necessitating the use of more flexible models. In this context, zero-inflated Poisson models have emerged as an effective statistical solution for dealing with this type of data. We choose the zero-inflated negative binomial (ZINB) distribution over the zero-inflated Poisson (ZIP) distribution. If var > mean, there is significant overdispersion. Scientific productivity, often measured by the number of publications in peer-reviewed journals, is a key metric for assessing academic performance, obtaining research funding, and advancing researchers' careers. Understanding and predicting the factors influencing researcher productivity provides valuable insights for institutions, policymakers, and researchers themselves. However, modeling scientific productivity faces methodological challenges due to the nature of publication data, which are often characterized by a large number of zeros (i.e., researchers who have not published any articles within a given time period) and excessive dispersion (i.e., the variance in the data is greater than the average). Researchers hypothesize that zeros in the data arise from two different sources: the first represents the complete absence of an event, and the second represents conventional conditions that can generate both zero and non-zero values. Recent years have seen significant advances in machine learning algorithms, which have proven their ability to handle complex, nonlinear data without the need for strict distributional assumptions. These algorithms include boosted trees, random forests, and artificial neural networks. Despite the growing interest in modeling scientific productivity, most current research relies on traditional statistical techniques, such as Poisson regression or negative binomial regression. However, these models often fail to address key features common in real-world academic productivity data: an excess of zeros and overdispersion (variance greater than the mean). Furthermore, while zero-inflation (ZIP/ZINB) models improve their handling of these characteristics, they are rarely systematically combined or compared with modern machine learning methods, which are increasingly used for forecasting in complex and high-dimensional data environments. However, machine learning models often sacrifice interpretability for predictive power. While statistical models are more interpretable, they may be less accurate in prediction.The predictive and interpretive power of zero-inflation regression models is evaluated compared to machine learning algorithms.

This study examines the integration of statistical and machine learning approaches to modeling scientific publication counts, with the goal of informing future research evaluation strategies that are both accurate and interpretable.Samuel M. Mwalili et al (2008)[2] The DMFT index is a popular measure in dental caries research, but its distribution is complex. Liuzzi and Thompson1 show that the zero-inflated negative binomial (ZINB) distribution gives an excellent fit. They fit this distribution to the DMFT index for 4,468 seven-year-old children. Alessandro Araldi et al (2020)[3]

comparison Two advanced modelling process: Penalised Count Regression and Machine Learning methods are implemented on the identical voluminous data of street-based observations of the urban form of the French Riviera. Abdullahi A. Ibrahim et al(2020)[4] . The fact that we had zero-inflated count data made us use k-inflated count models and machine learning algorithms to travel data to compare predictions, model fit, and find highly significant covariates. The analysis demonstrates that the models with k-inflation are a good fit, whereas machine learning algorithms can make better predictions. Monika Arora et al (2021)[5]We discussed the performance of some machine learning methods, using the catBoost classification algorithm as a case study, in both loan approval and employee promotion. We compared the algorithm's performance with other classifiers. Intesar N. El-Saeiti & Gadir Alomair(2023)[6] . They compared the four different models: zero-infested Poisson (ZIP), barrier Poisson (HurP), zero-infested negative binomial (ZINB), and barrier negative binomial (HurNB) based on balanced and unbalanced data conditions. We find that the ZIP model provides the most satisfactory fit to the claims frequency data in balanced and unbalanced cases of data. Yuwen Gu(2024)[7]. He analyzed the dispersion modeling issue on Tweedy application in the zero-inflated Tweedy model and modeled the probability of the zero case under the generalized expectation-maximization algorithm. Also, he determined the average, variation and probability of the zero instance with the help of decision tree-based gradient boosting. It has been substituted with a zero-inflated binary negative model. Here, the aggregated claims volume which is very much zero-infested is modeled by a compound negative binary model and a gamma mixture. The new model is capable of managing the situations whereby the number of claims lodged per policy has a large variation. Nonetheless, it has a higher computational cost than this model. Bonelwa Sidumoet al(2024)[8] Count data of the environment is often overdispersed, which is investigated in this study. We examine the most widely used count regression models that are Poisson, negative binomial, zero-inflated Poisson, and zero-infused negative binomial and evaluate how they address such data. These parametric models are then compared to the performance of four mainstream machine-learning regression models namely random forests, support vectors, k-nearest neighbors and artificial neural networks. Also, we explore the higher-order extensions of the Tweedy regression model to solve its short-comings in the modelling of aggregate claims of various forms of insurance, such as auto, health, and liability. Banghee So et al(2024) [9]We used Zero-Inflated Tweedie Boosted Trees and CatBoost, The effectiveness of our improved Tweedy model is demonstrated by applying an insurance telematics dataset. Our modeling results show a significant improvement in the model's performance, demonstrating its ability to provide more accurate predictions suitable for insurance claims analysis. Banghee So(2024)[10] Process insurance claims data, including zero-inflined telematics data, using boosting algorithms in order to train the claims frequency models. Three well-known libraries of gradient boosters were tested XGBoost, LightGBM, and CatBoost and CatBoost turned out to be the most effective in the creation of automated claims frequency models on the basis of predictive performance. In the study of Marko Martinovi et al(2025) [11]. have multiple ma- chine learning models. Some of the methods used were ran- dom forests, gradient boosting structures, support vector machines, neural networks, and logistic regression, validation technologies. The findings demonstrated that the tree-based boosting algorithms always perform well in terms of accuracy, precision in comparison to other models.

This research consists of three main sections, distributed as follows:

Section One: Includes a general introduction to the research, the problem, its importance, objectives, and limitations. It also defines basic terms, previous studies

related to count data, excessive zero, statistical models such as Poisson regression and negative marginal regression, and machine learning models, such as neural networks, random forests, support vectors, and therein. Section Two: Briefing the theoretical framework which will consist of two parts: the first part will be briefing on the statistical models which include Poisson regression, negative binomial regression, the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models and the second part will be a briefing on machine learning algorithms which include neural networks, support vector machines, random forests and boosting like XGBoost. The third section dwells upon model evaluation criteria. Part One: Practical Application includes the description of the data that was used, the methodology that will be used to divide the dataset into the training and testing part, the statistical models and algorithms, the analytical tools that will be implemented in R, and the performance indices.The second part covers the results of the statistical analysis and comparisons between the models, with a detailed explanation of the performance of each method in terms of predictive accuracy and fit criteria such as RMSE, MAE, AIC, LogLik, and others.The third part presents the conclusions and recommendations reached based on the results, presents a set of scientific recommendations for future studies, and suggests the use of advanced models in similar situations..

Theoretical Aspect:

Overdispersed Count Data:

Count data are data that represent the number of times an event occurs, and are often represented using the Poisson distribution and the basic assumption that mean = variance. Count data that are over-dispersed is known as "over-dispersion." This is count data in which the variance is larger than the mean. This contradicts the assumption of the traditional Poisson distribution, which is often not the case in real-world data, where the variance is much larger than the mean. Ignoring this over-dispersion leads to incorrect estimates and inaccurate benchmarks. The Pearson based dispersion statistic proposed by Hilbe [12] is used in assessing overdispersion of regular count models. If the ratio of the estimator.

$\hat{\varphi} = \frac{D}{\mathrm{n-p}}$ when D represents the deviance, and n -p (n is the number of observations and p is the number of parameters) represents the number of degrees of freedom exceeds one, evidence of overdispersion exists [13]. Interpreting the dispersion ratio:

When the ratio of dispersion is near to one, the data is well fitted by the Poisson model. A ratio value larger than one is necessary to show overdispersion and hence negative binomial or other such models can be a superior fit. The dispersion ratios that are significantly lower than one imply underdispersion. A $p < 0.05$ indicates overdispersion or underdispersion (the former is more common)[14].

Poisson regression (PR)

The Poisson distribution is the simplest form of distribution that is used to model the count data with a probability density function [15].

$$\mathrm{f(y;\ \lambda)} = \frac{exp^{-\lambda}\ \lambda^{y}}{\mathrm{y!}} \text{ ..... (1)}$$

This model is in the family (1), and thus Poisson regression is a particular case in the generalized linear model. The canonical link is $g(\lambda) = log(\lambda)$ with the consequence that the mean and linear predictor are associated log-linearly. The Poisson model variance is the same as the mean hence dispersion is fixed at $\varphi = 1$

and the variance function is $V(\lambda) = \lambda$.

Using the logarithmic link function ("log" in r), the general form of a Poisson glm is

$$\begin{cases} y & \sim \text{Pois}(\mu) \\ \log \mu & = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \end{cases}$$

The systematic component of (10.1) can be written as

$$\begin{aligned} \mu &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \\ &= \exp \beta_0 \times (\exp \beta_1)^{x_1} \times (\exp \beta_2)^{x_2} \times \cdots \times (\exp \beta_p)^{x_p}. \end{aligned} \quad \text{.....(2)}$$

This shows that the impact of each explanatory variable is multiplicative. Increasing x_j by one increases μ by factor of $\exp(\beta_j)$. If $\beta_j = 0$ then $\exp(\beta_j) = 1$ and μ is not related to x_j. If $\beta_j > 0$ then μ increases if x_j increases; if $\beta_j < 0$ then μ decreases if x_j increases.[16]

Negative binomial regression (NBR)

The negative binomial distribution is one of the most widely used discrete probability models, allowing for deviations from the Poisson model of equal mean and variance. More specifically, the negative binomial distribution exhibits overdispersion of data compared to the Poisson distribution..[17]

When the count response is overdispersed (variance > mean), the negative binomial distribution is the best choice. The relative likelihood coefficient for the response variable Y from the negative binomial distribution is:[5]

An appropriate model to use to model overdispersed count data is a negative binomial (NB) distribution of the conditional model of yi|x i, which may be thought to be a gamma mixture of Poisional distributions. The probability density function associated with it is determined by one parameter, which is: [15].

$$f(y; \mu, \theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta) \cdot y!} \cdot \frac{\mu^y \cdot \theta^\theta}{(\mu+\theta)^{y+\theta}}, \quad \text{..... (3)}$$

with mean μ and shape parameter $\boldsymbol{\theta}$; $\boldsymbol{\Gamma}(\cdot)$ is the gamma function. For every fixed $\boldsymbol{\theta}$, this is of type (1) and thus is another special case of the GLM framework. It also has φ = 1 but with variance function $\mathbf{V}(\mu) = \mu + \frac{\mu^2}{\boldsymbol{\theta}}$. [15]**.**

2.4 Zero-Inflation Model: Traditional Poisson and negative-binomial (NB) models are often biased to use on strongly dispersed data, where there are high and very large numbers of zero values. To this end, scholars have proposed the zero- rate model or zero-inflation model to deal with this problem in a proper manner [18][19], These processes represent a two-state model, where accident data are separated into two different states, one of which is a safe state (characterised by zero counts) and the other an unsafe state (characterised by non-zero counts). The sites that portray a very low or insignificant risk of accidents are categorized in the safe state and those that follow Poisson or the negative binomial (NB) distributions are categorized within the unsafe state. As such, the suitable zero-inflated probability models would be the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models [19].

Zero-Inflated Poisson (ZIP) Models:

The zero-inflation Poisson (ZIP) model has proven to be very efficient in handling data containing a large number of zero values, especially in cases where the variance is relatively limited. Zero-inflated Poisson (ZIP) models to accommodate the zero-inflated count data were introduced first by Lambert (1992).[20] Suppose the sample size is n, and counts of observations are independent, and the values are y i. In zero-inflated Poisson (ZIP) regression, the variable counts, y i, are the counts with probability p i, but otherwise distributed as a mixture of two other component distributions. As may be deduced, the counts are distributed according to Poisson distribution with mean of \mu i and probability of occurrence of 1-p i. The ZIP model can thus be seen as a mixture of two component distributions It can be derived that

We consider a zero-inflated Poisson regression model in which the response variable has the distribution model in which the response variable $Y_i (i = 1, \dots, n)$ has the distribution. [21]

$$\Pr(Y_i = y_i) = \begin{cases} \varphi_i + (1 - \psi_i)\exp(-\lambda_i), & y_i = 0 \\ (1 - \varphi_i)\exp(-\lambda_i)\lambda_i^{Z_i}/y_i!, & y_i > 0 \dots \dots . (4) \end{cases}$$

where the parameter λ_i and φ_i satisfy $\log(\lambda_i) = \sum_{j=1}^{k} x_{ij}\beta_j$ and $0 < \varphi_i < 1$. The mean and the variance of the distribution are $E(Y_i) = (1 - \varphi_i)\lambda_i$ and $\mathrm{Var}(Y_i) = (1 - \varphi_i)\lambda_i(1 + \varphi_i\lambda_i)$ Both the mean λi of the underlying Poisson distribution and the mixture parameter pi (also referred to as 'zero–inflation parameter') can depend on vectors of covariates.[22]

Alternative Models: ZINB:

When overdispersion is high, the ZIP model may become insufficient, making models like ZINB more appropriate due to their ability to accommodate higher variance. When overdispersion is greater than ZIP's ability to handle, the ZINB (zero-inflated negative bound) model is used, which allows for greater variance.The two most common models for count data are the Poisson model and the binomial model. The Poisson distribution has only one parameter and equal variance, a condition called equal dispersion. However, in reality, this is often not the case. For example, when the variance is less than the expected value of the response variable, this is known as underdispersion. Another situation may occur when the variance is greater than the expected value of the response variable, known as overdispersion. The reason behind this scenario is that the proportion of zero observations in the response variable (p) is too large and can be viewed as excessive or more than zero. The negative binomial distribution, however, has two parameters. In turn, the negative binomial regression model is more flexible in comparison to the Poisson regression model. Negative binomial regression model can be used instead of Poisson regression model where assumption of equal mean and variance is violated especially where the data under investigation has a high level of dispersion. [21] [23].Greene (1994) A generalized form of the negative binomial model of excess zero count data, the zero-inflated negative binomial (ZINB) has been proposed. Overdispersion, in that case, is caused by the heterogeneity of data, as well as by the excess of [24]zeros, The ZINB regression model tend to be more suitable as compared to the ZIP. In the ZINB regression model, the probability mass function of the response variable y i is a random non-negative integer-valued variable and it is assumed that yi = 0 is observed at a much higher frequency than can be modeled by the usual model. Thus, the regression model is defined as

$$P(Y = y_i \mid x_i, z_i) = \begin{cases} \varphi_1 + (1-\varphi_i)f(0;\theta_i), & y_i = 0, \dots\dots\dots (5) \\ (1-\varphi_i)f(y_i;\theta_i), & y_i > 0, \end{cases}$$

Where $f(y_i;\theta_i), y_i = 0,1,2,\dots$ is the pdf of Y_i and $0 < \varphi_i < 1$. Furthermore, the function $\varphi_i = \varphi_i(z_i)$ satisfies $\text{logit}(\varphi_i) = \log(\varphi_i[1-\varphi_i]^{-1}) = \sum_{j=1}^{m} z_{ij}\delta_j$ where $z_i = (z_{i1} = 1, z_{i2}, \dots, z_{im})$ is the i-th row of covariate matrix Z and $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ are unknown Column vector m- dimensions of parameters. In this setup, the non-negative function φ_i is modeled Via login link function. This function is linear and other appropriate link functions that allow it $\varphi_i$to be negative can be used. [22][21]

Machine Learning Algorithms:

Machine learning algorithms are powerful tools for building predictive models without relying on strict statistical assumptions. Some of the most prominent algorithms used with digital data include:

Random Forests:

These rely on ensemble analysis of a large number of decision trees and improve performance through averaging. Random forests were introduced by Leo Breiman [25][26]. who was inspired by the earlier work of Amit and Geiman [27]. Random forests are an extension of Breiman's bagging idea [28] and were worked out as a rival to boosting. A categorical response variable may be analyzed using the random forests, this is known as in [25][26]. as a classification, or a continuous response, termed as regression. On the same note, predictor variables may be categorical or continuous. Computationally, random forests are attractive because of the following: • they perform both regression and (multi-class) classification: they take this feature naturally, and do not require any additional tuning parameters; (they are relatively fast to train and predict); they can be easily parallelized: (they can be used directly in high-dimensional problems); they will give an estimate of generalization error; [29] [30]

Algorithm 2 Random Forests $(xi, 1, \dots, xi, p)T. For\ j = 1\ to\ J: Let\ D = \{(x1, y1), \dots, (xN, yN)\}$ denote the training data, with xi = $(xi, 1, \dots, xi, p)^T$ For j = 1 to J:

1. Take a bootstrap sample Dj of size N from D.

2. Using the bootstrap sample Dj as the training data, fit a tree using binary recursive partitioning:

a. Start by considering all the observations as one node.

b. The next steps are then applied recursively to all the nodes which have not yet been split until the required stopping criterion is reached: i. Sample m predictors out of the p predictors randomly.

ii. Find the best binary split among all binary splits on the m predictors from step i.

iii. Split the node into two descendant nodes using the split from step ii.

To make a prediction at a new point x, •

$\hat{f}(x) = \frac{1}{J} \sum_{j}^{J} \hat{h}\, j(x) \dots\dots (6)$ for regression

where $\hat{h}\, j(x)$ is the prediction of the response variable at x using the jth tree[30].

Boosting (XGBoost):

Boosting, as with bagging, but trees are planted in series: the next tree is planted utilizing data of trees that have already been planted, and the aim is to minimize the error of the former models.[31] Boosting resembles the ensemble approach, but trees are trained in serial order that is, the new tree is trained with the knowledge of the already trained trees, with the aim of minimizing the error of the predecessor models[28] in a regression tree model, the process is as follows:

Add a decision tree to the model that uses the residual errors of the model as the outcome variable. Call this new decision tree, with the lambda shrinkage parameter, to the fitting function to replace the residuals. Lambda is a small positive number, normally it lies between 0.01 and 0.001 [32] The result of this method is the successive, incremental improvement of the fitted model, which produces a high-performance model. Boosting can have different tuning parameters such as:

Number of trees

Lambda shrinkage parameter

Number of splits in each tree.

There are various forms of boosting, including Adaboost, gradient boosting, and stochastic gradient boosting.These, such as XGBoost and CatBoost, create a series of weak models that are gradually boosted to improve performance.[33]

Neural Networks:

They are computing systems inspired by the biology of neural networks that make up the brains of humans and animals. These systems gradually improve their performance in prediction and classification by "learning" from previous iterations of computations and making adjustments. Neural networks consist of a collection of units known as "neurons," and each connection between them can convey information. These connections, in turn, can have a weight assigned to them as the model approaches a solution, which may increase or decrease the influence of a particular neuron on the outcome. In effect, neural networks resemble a non-parametric regression model, allowing for more robust functional performance[34][35] [36]. As with biological neural networks, an artificial neural network also consists of neurons connected to each other by weights that play the same role as synaptic strength.[37]

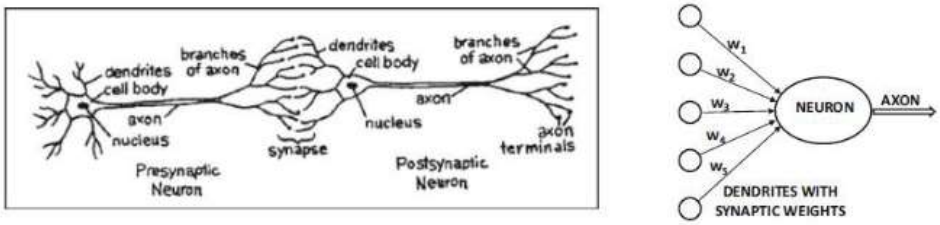


Figure 1. Biological Neural Network versus Artificial Neural Network. [37].

If we analyze the basic structure of an artificial neural network (ANN), we will notice some common elements in any neural model [38][37]. There are four main components in a neural model:

1. The initial element is the synapse that is denoted by a weight. The specific, adjustable weight is linked to each input and alters the strength of the input with the desired result.

2. The second element is referred to as the adder and is linked to an operator that adds all the input signals by the synaptic strength.

3. The third element, the bias, is a part of a linear equation. It is a special neuron which holds another value that is used to fit the data to the model in the most appropriate manner.

4. Lastly, there is the activation function which defines the amplitude of the neurons with regard to a given range as illustrated in the figure below [37].

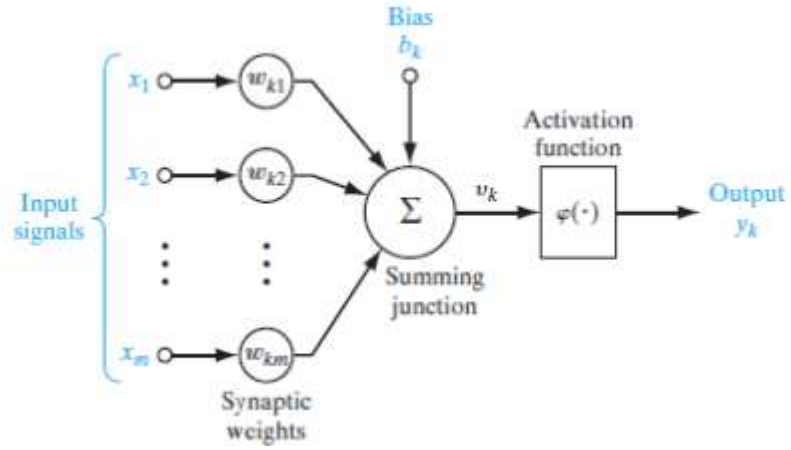


Figure 2. Key elements of an Artificial Neural Network[37].

Support Vector Machines (SVMs):[39]

Support vector machines (SVMs) are a state-of-the-art supervised machine learning technique [40][39]. SVM models are closely related to classical multi-layer perceptron neural networks. The idea behind SVMs is that of a margin the side of a hyperplane between two data classes. It is proved that the maximum desired distance between the separating hyperplane and instances on one side as well as the other, that is, the maximization of the margin, reduces the maximum anticipated generalization error [41]

. By using a different loss function called the ϵ-insensitive loss function $\|y - f(x)\|_\varepsilon = max\{0, \|y - f(x)\| - \epsilon\}$, SVMs can also perform regression. This loss function ignores errors that are smaller than a certain threshold $\epsilon > 0$ thus creating a tube around the true output.[42] The primal becomes:

$$\begin{aligned} &\text{minimize} && t(\mathbf{w}, \xi, \rho) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \rho + \tfrac{1}{m\nu}\textstyle\sum_{i=1}^{m} \xi_i \\ &\text{subject to} && \langle \Phi(x_i), \mathbf{w}\rangle + b \geq \rho - \xi_i \ (i = 1, \ldots, m) \\ & && \xi_i \geq 0 \ (i = 1, \ldots, m) \end{aligned} \quad \ldots\ldots\ldots\ldots (7)$$

Support vector machine regression can be evaluated by approximating the scale parameter of a Laplacian distribution on the residual ζ = y – f(x), where f(x) represents the value of the estimated decision function. [43]

Model Evaluation Criteria:

To selecting the model, evaluate the precision of the estimates to determine its accuracy. AIC stands for (Akaike's Information Criteria), a metric developped by the Japanese Statistician, Hirotugu Akaike, 1970. defined as $AIC = 2k - 2\log(\ell(y; \hat{\mu}))$, where k is the number of parameters included in the model and ℓ is the likelihood of the fitted model defined as above. [44] The fundamental concept of AIC is to punish

the addition of more variables into a model. It introduces a penalty which amplifies the error upon the addition of more terms. The smaller the AIC the more desirable the model[44].

Loglik performance evaluation of the postulated the Regularization regression metho The likelihood approach is contrasted in its performance that is based only on the evidence by the observed data (excluding the most extreme values) and the likelihood function. The likelihood ratio is represented by the ratio of the heights below the likelihood function in accordance with the values of the tested hypothesis. In order to have a linear measure of evidence, the natural logarithm of this is calculated to achieve the logarithmic likelihood ratio [45].

Likelihood ratio tests (LRTs) have been used to compare two models. nested The form of the test is suggested by its name:[46]

$$LRT = -2\, log_e \frac{\zeta_s(\hat{\theta})}{\zeta_g(\hat{\theta})} \quad \text{.....(8)}$$

ds and comparing ability with other forms of existing penalized algorithms is most commonly judged by the variables selection and evaluation criteria of prediction error rates and the quality of the model used in regression analysis.MAE (Mean absolute error), RMSE (Root mean squared), and R-Squared are most commonly used measures of assessment of the rates of the prediction errors and the quality of the model that is used during regression analysis.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \qquad (9)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2} \qquad (10)$$

- **R-squared** (Coefficient of determination) is the coefficient of the exact fit of the values to actual values. The interpretation of the value between 0 and 1 was in terms of percentages. The greater the value the better is the model.

The above metrics can be expressed,

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \qquad (11)$$

Where,
$\hat{y}$ – predicted value of y $\bar{y}$ – mean value of y

3.1.Real application

The "art" variable is the primary outcome of the count, representing the number of academic publications. https://cran.r-project.org/web/packages/pscl/pscl.pdf The dataset contains a large number of zeros in "art" (many researchers have not published any articles [47][48].

art Numeric (Count) Number of articles published by the biochemist – the response variable

fem Factor (Binary: 0/1) Gender: 1 = Female, 0 = Male

mar Factor (Binary: 0/1) Marital status: 1 = Married, 0 = Single

kid5 Numeric (Integer) Number of children under age 5

phd Numeric (Continuous) Years since PhD was awarded

ment Numeric (Integer) Number of articles published by the mentor (PhD advisor)

We split the dataset into two parts with 80 percent of it being used as training and the remaining 20 percent used as testing. The training sub-set was utilized to fit the models in addition to optimizing the hyperparameters, and the testing sub-set was used to evaluate the predictive performance of each model in addition to the generalizability.

Apart from machine learning algorithms, including neural networks, support vector machines, and random forests, apply statistical models, including Poisson regression and negative binomial regression.

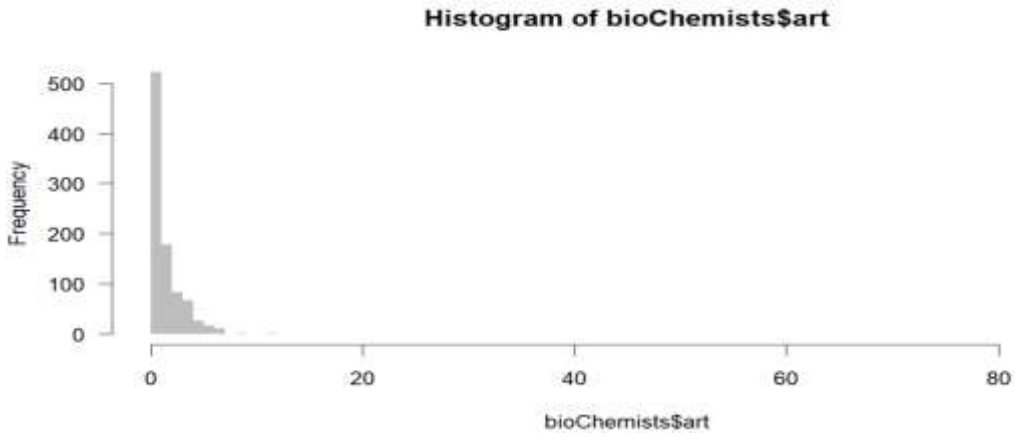


Figure 3. Histogram of Academic Productivity (art)

Table 1. Evaluation of Zero-Inflated Regression Models and machine learning methods using dispersion, AIC, BIC, log-likelihood, RMSE, MAE, and pseudo R^2

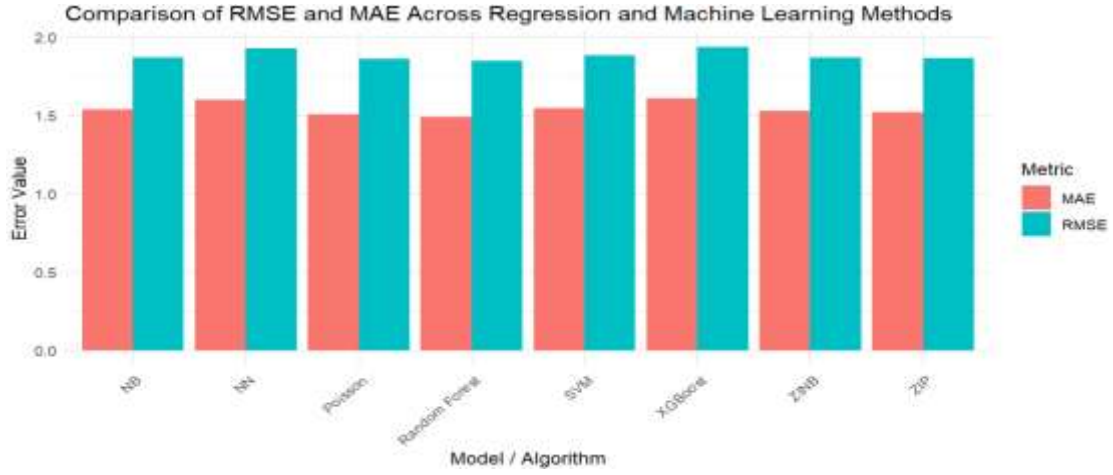


Figure 4. representing the performance comparison of Zero-Inflated Regression Models and machine learning methods using RMSE and MSE.

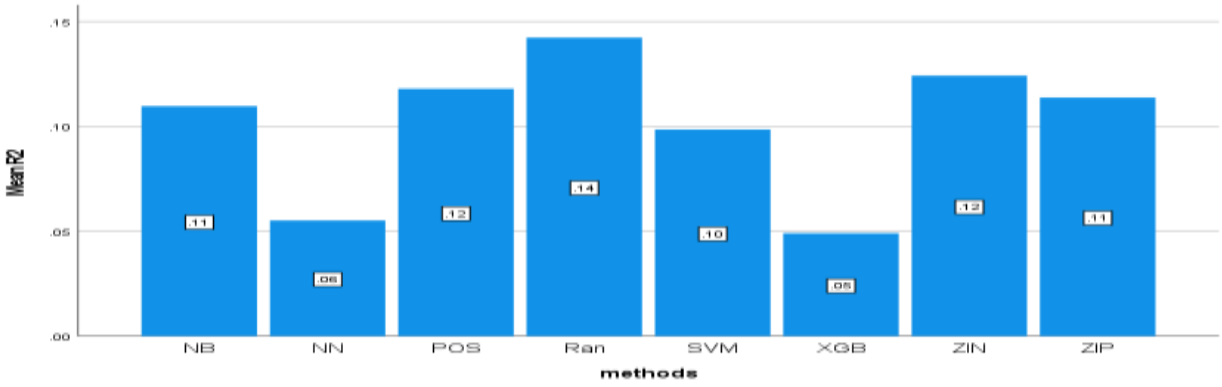


Figure 5. R^2 Values Zero-Inflated Regression Models and machine learning methods.

Interpretation of Results:

Poisson Regression (POS) AIC = 2651.62, BIC=2679.19,LogLik = -1319.81 Regression (POS) – Dispersion = 1.82, greater than 1. There is significant overdispersion in the data ,Dispersion =1.82Interpretation: The model suffers from significant overdispersion (dispersion > 1), making it an inadequate fit to the data. The AIC and LogLik are higher than the other models, indicating poor statistical fit. Negative Binomial Regression (NB) AIC = 2506.57 , BIC=2538.74 LogLik = -1246.28 , Dispersion = 1.11.The model handles overdispersion well, showing strong performance in terms of AIC and LogLik. It is considered best suited if overdispersion is not a major problem. Zero-Inflated Poisson (ZIP): AIC = 2587.00, BIC=2643.09,LogLik = -1281.97 .It handles excessive zero well but does not handle overdispersion, and its performance is weaker than NB and ZINB. Zero-Inflated.NB: AIC = 2513.58, BIC=2545.75,LogLik = -1249.79.This model combines the advantages of ZIP and NB, and addresses both over-dispersion and over-zero dispersion.Making it an excellent choice in both theory and practice. Random Forest achieved the best predictive performance (lowest RMSE ≈ 1.847) SVM achieved average results (RMSE ≈ 1.884): Neural Networks and XGBoost performed relatively poorly (RMSE > 1.93):

Conclusions and Recommendations:

Conclusions:

Random Forest achieved the best predictive performance (lowest RMSE), outperforming other machine learning algorithms without the need for complex parameter tuning. It also efficiently handles interactions between variables and nonlinear distributions. SVM achieved average results (RMSE) and performs relatively well but is sensitive to tuning parameters. Neural networks and XGBoost performed relatively poorly (RMSE > 1.93), and despite their theoretical power, they did not demonstrate clear effectiveness in this case without in-depth optimization.

Poisson regression alone is not recommended here. Poisson regression (POS) - dispersion = 1.82, greater than 1. There is significant overdispersion in the data. This means that the Poisson model is not suitable because it assumes equal variance and mean (equal dispersion).Statistical models such as Poisson/NB/ZIP/ZINB were competitive, despite their simplicity, and provided results very close to those of machine learning algorithms. These models are suitable for data containing many zero values and overdispersion. Use the Random Forest model as your baseline model if

you want the best predictive accuracy without the need for complex tuning.We use the ZINB or ZIP model if there are many zero values and a large amount of dispersion, and if we are interested in explaining the variables with more than just accuracy.If the data contain both zero overdispersion and overdispersion, we use the ZINB model. If the data contain only overdispersion (without significant zero overdispersion), we use the negative binomial model.If there is no overdispersion or zero overdispersion (rarely), we can use the Poisson model.

Recommendations:

Bayesian zero-inflated regression models are recommended for future studies, given their ability to handle uncertainty and random variance, especially in non-uniformly distributed count data. Although XGBoost is used, other algorithms such as LightGBM and CatBoost have demonstrated superior performance in many applications and are promising alternatives for dealing with imbalanced count data. Zero-inflated mixed models or hierarchical models can be applied when data are hierarchical. The performance of some machine learning algorithms (such as XGBoost and NN) can be significantly improved by using grid search or layered Bayesian optimization (BIO) cross-validation (STR).

References

[1] W. Shankar, V.; Mannering, F.; Barfield, *Effect of roadway geometrics and environmental factors on rural freeway accident frequencies.* 1995.

[2] S. M. Mwalili, E. Lesaffre, and D. Declerck, "The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research," *Statistical Methods in Medical Research*, vol. 17, no. 2, pp. 123–139, 2008, doi: 10.1177/0962280206071840.

[3] A. Araldi, A. Venerandi, and G. Fusco, "Count Regression and Machine Learning Approach for Zero-Inflated Over-Dispersed Count Data. Application to Micro-Retail Distribution and Urban Form," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12252 LNCS, pp. 550–565, 2020, doi: 10.1007/978-3-030-58811-3_40.

[4] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 738–748, 2020, doi: 10.14569/IJACSA.2020.0111190.

[5] M. Arora, Y. Kalyani, and S. Shanker, "A comparative study on inflated and dispersed count data," *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021*, no. Data, pp. 29–38, 2021, doi: 10.5220/0010547700290038.

[6] I. N. El-Saeiti and G. Alomair, "Comparative Evaluation of Zero-Inflated and Hurdle Models for Balanced and Unbalanced Data: Performance Assessment and Model Fit Analysis," *European Journal of Science, Innovation and Technology*, vol. 3, no. 1, p. 207, 2023, [Online]. Available: www.ejsit-journal.com.

[7] Y. Gu, "Dispersion Modeling in Zero-inflated Tweedie Models with Applications to

Insurance Claim Data Analysis," pp. 1–24, 2024, [Online]. Available: https://arxiv.org/abs/2405.14990v1.

[8] B. Sidumo, E. Sonono, and I. Takaidza, "Count Regression and Machine Learning Techniques for Zero-Inflated Overdispersed Count Data: Application to Ecological Data," *Annals of Data Science*, vol. 11, no. 3, pp. 803–817, 2024, doi: 10.1007/s40745-023-00464-6.

[9] B. So and E. A. Valdez, "Zero-Inflated Tweedie Boosted Trees with CatBoost for Insurance Loss Analytics," no. June, 2024, doi: 10.1016/j.asoc.2025.113226.

[10] B. So, "Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM," *Scandinavian Actuarial Journal*, no. June, 2024, doi: 10.1080/03461238.2024.2365390.

[11] M. Martinović, K. Dokic, and D. Pudić, "Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach," *Applied Sciences (Switzerland)*, vol. 15, no. 7, pp. 1–44, 2025, doi: 10.3390/app15073636.

[12] Hilbe JM, "Negative binomial regression.," *Cambridge University Press, Cambridge. https://doi. org/CBO9780511973420*, vol. 10, no. 1017, 2011.

[13] W. N. et al Zuur AF, Ieno EN, "Mixed effects models and extensions," *in ecology with R. Springer, Berlin*, 2009.

[14] S. Nakagawa, P. C. D. Johnson, and H. Schielzeth, "The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of the Royal Society Interface*, vol. 14, no. 134, 2017, doi: 10.1098/rsif.2017.0213.

[15] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in R," *Journal of Statistical Software*, vol. 27, no. 8, pp. 1–25, 2008, doi: 10.18637/jss.v027.i08.

[16] P. K. Dunn and G. K. Smyth, *Generalized Linear Models With Examples in R, https://link.springer.com/book/10.1007/978-1-4419-0118-7*. 2018.

[17] D. Cahoy, E. Di Nardo, and F. Polito, "Flexible models for overdispersed and underdispersed count data," *Statistical Papers*, vol. 62, no. 6, pp. 2969–2990, 2021, doi: 10.1007/s00362-021-01222-7.

[18] . Miaou, S.P., Song, J.J., "Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence.," *Accident Analysis and Prevention,* vol. 37, no. 4, pp. 699–720., 2005.

[19] PRATHYUSHA VANGALA, "NEGATIVE BINOMIAL-GENERALIZED EXPONENTIAL DISTRIBUTION: GENERALIZED LINEAR MODEL AND ITS APPLICATIONS," 2015.

[20] T. Loeys, B. Moerkerke, O. de Smet, and A. Buysse, "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression," *British Journal of Mathematical and Statistical Psychology*, vol. 65, no. 1, pp. 163–180, 2012, doi: 10.1111/j.2044-8317.2011.02031.x.

[21] Seyed Ehsan Saffari and Robiah Adnan, "Zero-Inflated Negative Binomial Regression Model with Right Censoring Count Data," *Journal of Materials Science and Engineering B*, vol. 1, no. 9, 2011, doi: 10.17265/2161-6221/2011.09.020.

[22] D. Kokotsaki, V. Menzies, and A. Wiggins, "Durham Research Online Woodlands," *Critical Studies on Security*, vol. 2, no. 2, pp. 210–222, 2014, doi: 10.1002/bimj.200100000.

[23] R. Fitriani, L. N. Chrisdiana, and A. Efendi, "Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data," *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052025.

[24] L. . K. Rahayu, *Overdispersi Pada Regresi Poisson dan Zero Inflated Poisson untuk Beberapa Karakteristk Data.*, : : Bogor Se. 2014.

[25] L. Breiman, "Random Forests.," *Machine Learning*, vol. 45, no. 1, pp. 5–32.

[26] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, "Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors," *J. Mater. Chem. C*, vol. 3, no. May, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[27] D. Amit, Y., Geman, "Shape quantization and recognition with randomized trees.," *Neural Computation 9(7) pp.*, vol. 9, no. 7, pp. 1545–1588, 1997.

[28] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 2001.

[29] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Machine Learning*, no. January, 2012, doi: 10.1007/978-1-4419-9326-7.

[30] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Machine Learning*, no. February 2014, 2012, doi: 10.1007/978-1-4419-9326-7.

[31] R. Hastie, T., James, G., Witten, D., & Tibshirani, *An introduction to statistical learning*. 2013.

[32] and R. T. James, Gareth, Daniela Witten, Trevor Hastie, *An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.* 2014.

[33] A. Kassambara, "Machine Learning Essentials. Practical Guide in R.," *Sthda*. p. 210, 2017, [Online]. Available: https://www.m-culture.go.th/mculture_th/download/king9/Glossary_about_HM_King_Bhumibol_Adulyadej's_Funeral.pdf.

[34] M. Warner, B. & Misra, "Understanding Neural Networks as Statistical Tools.," *Am Stat*, vol. 50, no. 4, pp. 284–293, 1996.

[35] B. Warner, "Understanding Neural Networks as Statistical Tools 1," 1996.

[36] WILLIAM ARTHUR LINDBLAD, "A COMPARISON OF THE PREDICTIVE ACCURACY OF LASSO REGRESSION AND NEURAL NETWORK MODELS, INTERNALLY AND

EXTERNALLY VALIDATED USING SIMULATED DATA," 2017.

[37] L. M. Figueiredo, "An Overview of the main Machine Learning Models From Theory to Algorithms," 2020.

[38] S. Haykin, *Neural Networks and Learning Machines Pearson Education.* 2009.

[39] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.

[40] V. N. Vapnik, "The Nature of Statistical Learning Theory.," *(2nd ed.). Springer Verlag. Pp. Retrieved from website: https://www.andrew.cmu.edu/user/kk3n/simplicity/vapnik2 000.pdf*, pp. 1 – 20, 1995.

[41] S. B. Kotsiantis, "Supervised Machine Learning," *A Review of Classification Techniques. Informatica . Retrieved from IJS website: http://wen.ijs.si/ojs-2.4.3/index.php/informatica/article/download/148/140.*, vol. 31, no. 10, pp. 249 – 268, 2007.

[42] A. Karatzoglou, D. Meyer, and K. Hornik, "Support Vector Algorithm in R," *Journal of Statistical Software*, vol. 15, no. 9, pp. 1–28, 2006.

[43] W. R. Lin CJ, ""," *Probabilistic Predictions for Support Vector Regression." URL http://www.csie.ntu.edu.tw/~cjlin/papers/svrprob.pdf*, 2004. .

[44] J. Lanford, T. Nykodym, A. Rao, and A. Wang, "Generalized Linear Modeling with H2O's R Package," no. February, 2021.

[45] P. Cahusac, "Log Likelihood Ratios for Common Statistical Tests Using the likelihoodR Package," *R Journal*, vol. 14, no. 3, pp. 203–212, 2022, doi: 10.32614/RJ-2022-051.

[46] W. E. Strawderman, "Likelihood Ratio Tests," *Wiley StatsRef: Statistics Reference Online*, pp. 1–2, 2014, doi: 10.1002/9781118445112.stat01645.

[47] A. Agresti, *Analyzing Contingency Tables*. 2019.

[48] Z. M. Simon Jackman, Alex Tahk, Achim Zeileis, Christina Maimone, James Fearon, *Package 'pscl': Political Science Computational Laboratory*. 2022.