Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 3, No. 2, Summer & Autumn 2023
Research paper

Modeling auto insurance frequency using K-means and mixture regression

Maryem Jaziri[1], **Afif Masmoudi**[2]

[1] University of Sfax , Probability and Statistics Laboratory
12011991maryem@gmail.com

[2] University of Sfax , Probability and Statistics Laboratory
afifmasmoudi@gmail.com

Abstract:
Given the importance of policyholder classification in helping to make a good decision in predicting optimal premiums for actuaries. This paper proposes, first, an optimal construction of policyholder classes. Second, Poisson-negative Binomial mixture regression model is proposed as an alternative to deal with the overdispersion of these classes. The proposed method is unique in that it takes Tunisian data and classifies the insured population based on the K-means approach which is an unsupervised machine learning algorithm. The choice of the model becomes extremely difficult due to the presence of zero mass in one of the classes and the significant degree of overdispersion. For this purpose, we proposed a mixture regression model that leads us to estimate the density of each class and to predict its probability distribution that allows us to understand the underlying properties of our data. In the learning phase, we estimate the values of the model parameters using the Expectation-Maximization algorithm. This allows us to determine the probability of occurrence of each new insured to create the most accurate classification. The goal of using mixed regression is to get as heterogeneous a classification as possible while having a better approximation. The proposed mixed regression model, which uses a number of factors, has been evaluated on different criteria, including mean square error, variance, chi-square test and accuracy. According to the experimental findings on several datasets, the approach can reach an overall accuracy of 80%. Then, the application on real Tunisian data shows the effectiveness of using the mixed regression model.

Keywords: Classification, K-means, Mixture regression, Overdispersion, MSE, Frequency.
Classification: G21, C13, C15, C61.

[1]Corresponding author
Received: 02/10/2023 Accepted: 13/01/2024
https://doi.org/10.22054/JMMF.2024.76043.1106

1 Introduction

Despite the large number of automobile accidents worldwide, their causes remain a difficult subject to determine in a definitive way due to the complex relationships between many contributing factors. The importance of the subject of modeling and forecasting traffic accidents has piqued the interest of many researchers,(S.Tang et al.[33]) . Under the umbrella terms "computer intelligence" and "data extraction," a variety of techniques and methodology have been used in the literature, (M.Lichman[20]), including neural networks, support vector machines, regression, decision trees, Bayesian networks, rules of association, clustering techniques, case-based reasoning, and ontologies. This study seeks to integrate several policyholder-specific factors through machine learning algorithms to arrive at an optimal prediction of the number of accidents for each new policyholder. Clustering techniques aim at discovering clusters of a set of models or data and are widely used in any discipline that involves the analysis of multivariate data (L.Breiman et al.[15],[17]). Their application in different fields is multiple and diverse. They can also be used in insurance in general. In order to test the usefulness and performance of the classes, we experiment with unsupervised K-means, taking into account the property of the machine learning-based technique. We adopt class selection to find a set of optimal classes especially when we have a well-seen zero mass in our dataset to improve the accuracy.

The determination of the occurrence frequency of a certain phenomenon, in particular the number of accidents in our case, is an information very much used by actuaries in the world of car insurance in order to determine the probability that an individual (insured) belongs to such a well-determined class. The stability of the classes has been evaluated on several random runs in terms of intra-cluster and inter-cluster to have an optimal number of classes.

In the modeling of accounting processes, such as the frequency of sinistres, two types of models are frequently used: the poisson regression model and the negative binomial regression model. There is a substantial body of literature on the application of these models: Shi and Valdez [25], Winkilmann [30], Greene [8], Yau et al. [31], Yang et al. [32].

The Poisson regression model has been frequently utilized in the insurance sector to model data on the quantity or frequency of claims. M.Aitkin et al [18] and Renshaw [1], for example, applied the Poisson model to two different sets of UK automotive claims data. The Poisson regression model has been considered practical and convenient for insurance practitioners; in addition to determining statistical inference and hypothesis testing using statistical theories. This model also allows the fitting procedure to be performed easily using any statistical software that includes an Iterative Weighted Least Squares (IWLS) regression routine.

When adopting a count model for assessing counting data, it is important to evaluate if there is at least overdispersion or an excess of zeros (e.g., Hinde and

Demétrio [11], Akantziliotou et al. [3], Sellers and Raim [26], or Del Castillo and Pérez-Casany [5]). When the variance (observed) exceeds the mean (expected) variance, overdispersion is most frequently present. This condition can also result from a zero-inflation sample (or an overabundance of zeros) or a long tail. Both measures are routinely applied to the Poisson distribution, and the negative binomial has lately been employed to model both types of count data sets.

Numerous writers investigated novel probability distributions based on the mixing mechanism. For instance, Simon [24] produced a negative binomial distribution by combining the mean of the Poisson distribution with the gamma distribution. Jerald [13] combined mixed Poisson regression with negative binomial regression. By combining the negative binomial distribution with the Lindley distribution, which has a thick tail and an alternative for modeling count data of insurance claims, which has a thick tail and a large value at zero, Zamani and Ismail [12],[34] established a new mixed negative binomial.

Because of their complex data structures, insurance loss data cannot always be well modeled by a single distribution. For univariate loss data, one alternative method in statistical analysis is to use a mixture of distributions. When the individual risk factors are available, a mixture of regressions is a natural extension to modeling the risk heterogeneity that the individual risk factors capture. Stephen and Richard [23] proposed the mixture model, which serves as the foundation for a Markov-switching regression model. Lindsay [16] and McLahlan and Peel [19] provide comprehensive reviews of mixing models. A mixture model, also known as a mixture of experts model in machine learning, is an ensemble learning technique that applies the idea of training experts on subtasks of a predictive modeling issue (Jiang and Tanner[29]). Recent research on insurance loss modeling has also focused on using mixture models to deal with complex data sets. To explain multivariate count data with extra zeros, Zhang et al.[35] proposed a multivariate zeroinflated hurdle model. Lee and Lin (2012) created a multivariate version of an Erlang mixture. To simulate insurance claim amounts, Lukasz et al.[28] developed a blend of neural networks with gamma loss. To simulate insurance claim amounts combining censored and shortened data, Roel et al.[22] used a mixture of Erlangs.

Fung et al. [6], Fung et al. [7], and Tseung et al. [27] proposed a class of so-called logit-weighted reduced mixture of experts (LRMoE) models for multivariate claim frequencies or severities distributions.

However, because there are a large number of insured without accident in the portfolio over an exercise period (one year), the number of zeros in the variable response (frequency of accident) is significant.
To address the importance of null values as well as the heterogeneity of the population, a Poisson-negative binomial mixture regression model has been proposed.

The interest of using a mixture regression is first of all to have a better approximation; moreover to have a classification as heterogeneous as possible.

Our empirical technique is novel in that it uses Tunisian data on the one hand,

and that we use the K-means machine learning method to classify the data, as well as a Poisson-NB mixture regression to forecast the optimal number of accidents for each new insured to predict an optimal premium. The performance of the proposed model was evaluated by several criteria such as, mean-squared-error (MSE), deviance, Chi-square test and accuracy.

There are two primary sections to this essay. The first section presents an indepth study of the methodology adopted in this work. First, we describe the K-means classification method. Then, the probabilistic or econometric counting models (Poisson, NB) for the distribution of accidents used as an estimation model are summarized. Finally, we presented a proposed new regression model (Poisson-NB mixture regression) for which the EM algorithm was used to learn its parameters.The second part is devoted to a real application on a Tunisian insurance company. The data is provided first, followed by some exploratory statistics. Second, the empirical findings are examined, which are divided into two sections: an interpretation of the explanatory variables' relevance and a comparison of our novel model to other regression models. The paper ends with conclusions that summarize the results obtained and discusses suggestions for further studies.

2 Methodology

In this section, we use the "K-means" classification algorithm in order to have an optimal number of classes that are as heterogeneous as possible between them. Then, we perform a generalized linear model (GLM) for each class. The presence of a mass at zeros in one of classes and the degree of overdispersion is quite high, which leads to a major problem of model choice. Here, we propose a mixture regression between the "Poisson" and "Negative Binomial" distributions. In the learning part, we use the (EM) algorithm to estimate the parameters model. Those, help us to predict the probability of appearance of each new insured in order to make an optimal classification.

The following diagram summarizes the structure of this work.

2.1 K-means

The objective of classification and unsupervised learning is to identify groups of observations with similar characteristics. From this classification, we aim to have individuals (the insured) in the same group come together and in different groups stand out as much as possible. Therefore, if a particular piece of data is a member of one cluster, it cannot be a member of another. There are several types of clustering, all of which have been well investigated in the literature. Because it is unsupervised and has a linear algorithmic complexity, we selected the K-means algorithm for this project. In this scenario, data will be assigned a suitable membership value, and the number of clusters will be established by running $K - means$ with various k

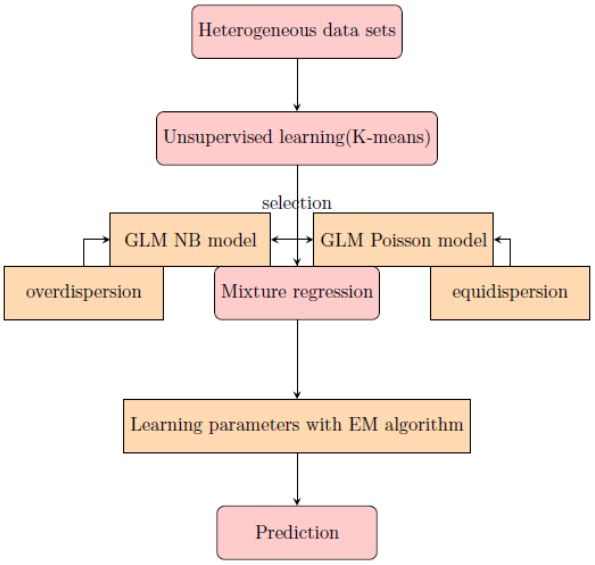

Figure 1: A graphic depicting the employed method.

values, computing the sum of squared errors of the various clusters ((SSE)), and determining the silouette Score (S).

The squared distances between the cluster is centroid and each member are added to form the SSE. As a result, we would to estimate a number of clusters k so that the distance between their centroids and the observations in the same cluster is as small as possible. We are talking about keeping the intra-class distance to a minimum. The silhouette score (S) is utilized to determine the optimal number of clusters for our dataset. In this case, we are attempting to optimize the inter-class distance between data points with cluster centers. In statistics and probability theory, the Konig-Huygens theorem is a remarkable identity between the variance and the mean. This theory allows to link the inter and intra inertia based on the following fundamental relation.

Total inertia= Inter - class inertia + Intra - class inertia

The optimal number of clusters K, according to Figure 2, is $K = 2$. The graphic on the left of Figure 2, which represents the intra-class, shows that when $K = 2$, the best minimization occurs. Furthermore, the inter-class silhouette Score suggests that $K = 2$, which is extremely close to 1, showing that the clusters are dense and well spaced, as well as being the curve's maximum point.

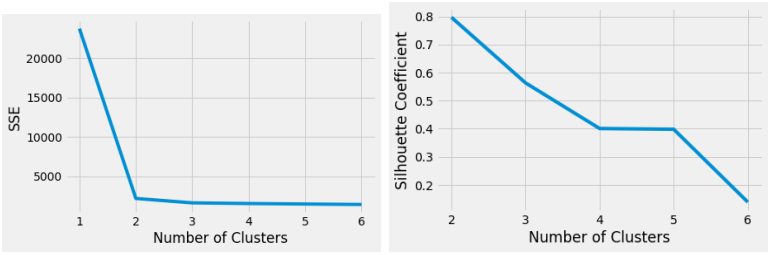

Figure 2: algorithmic k-means clustering of data.

2.2 GLM regression model

GLM extends the familiar linear regression models for quantitative data to include count and frequencies data, for which the assumption of normal errors is no longer reasonable.

GLM Poisson model:

Let Y be the random variable for claims number in the i^{th} policyholder. Assume that, Y follows a Poisson distribution with parameter λ. Its probability mass function is given by:

$$\mathbb{P}(Y = y) = Pois(y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \tag{1}$$

One property of Poisson regression model is mean-variance equality (or "equidispersion") conditional on explanatory variables :

$$\mathbb{E}(Y_i|X_i) = \text{Var}(Y_i|X_i) = \lambda_i = e^{X_i\beta}, i = 1, ..., n$$

We assume that the claims number Y_i of the i^{th} customer is a function of the covariates $X_i = [1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}]$, where x_i a $px1$ vector of explanatory variables, and β a $px1$ vector of regression parameters. If the conditional distribution of Y_i given X_i is a poisson distribution with parameter $\lambda_i = e^{X_i\beta}$ where $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8]$, then the conditional distribution of Y_i given X_i is Poisson. The coefficient regression parameter β will be estimated by the maximum likelihood method. The maximum likelihood estimator $\widehat{\beta}$ of β verify the following likelihood equations:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \lambda_i)x_{ij} = 0, j = 1, 2, ..., p \tag{2}$$

Since Eq(2) is also equal to the weighted least squares, the maximum likelihood estimates, $\widehat{\beta}$, may be solved by using the Iterative Weighted Least Squares (IWLS)

regression.
One can notice that this model allows the variance to exceed the mean. Moreover, the Poisson regression model can be regarded as a limiting model of the negative binomial regression model as $\nu = 1/\alpha$ approaches 0.

GLM Negative binomial model:

The negative binomial model is a generalization of the Poisson model for overdispersion in that it assumes that events occur contagiously and/or in a heterogeneous environment rather than simulating a series of isolated occurrences with a fixed expectation.

By adding a gamma noise variable with a mean of 1 and a scale parameter of nu, the Poisson distribution can be made more generic. The Poisson-gamma mixture (Negative binomial) distribution that results is constructed as: Consider that, the number Y of claims (accidents), given the parameter λ, is distributed according to the Poisson model with parameter λ, where λ denotes the different underlying mean risk of each policyholder to have an accident.

Let us assume that

$$Y|\lambda \sim \text{Poisson}(\lambda) \text{ and } \lambda \sim \Gamma(\mu, \alpha)$$

Where $\Gamma(\mu, \alpha)$ denotes the following Gamma density function:

$$\Gamma(\mu, \alpha) = \frac{\lambda^{\mu-1}\alpha^{\mu}\exp(-\alpha\lambda)}{\Gamma(\mu)} 1_{(0,+\infty)}(\lambda), \mu > 0, \alpha > 0,$$

mean $\mathbb{E}(\lambda) = \mu/\alpha$ and variance $Var(\lambda) = \mu/\alpha^2$.
Using the total probability formula, the negative binomial distribution of Y considered in this study has the following form:

$$\begin{aligned} NB(y|\mu,\alpha) &= \mathbb{P}(Y = y|\mu,\alpha) \\ &= \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \end{aligned} \quad (3)$$

A group of p regressor variables (also known as the explanatory variables x) are used in negative binomial regression to get the mean of Y. The expression relating these quantities is:

$$\mathbb{E}(Y_i|X_i) = \mu_i = e^{X_i\theta} \text{and } \text{Var}(Y_i|X_i) = e^{X_i\theta}(1 + \frac{1}{\alpha}e^{X_i\theta}),$$

In what follows, we assume that the claims number Y_i of the i^{th} customer is a function of the covariates $X_i = [1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}]$.
If the conditional distribution of Y_i given X_i is a poisson distribution with parameter $\mu_i = e^{X_i\theta}$, where $\theta = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8]^T$, then the conditional

distribution of Y_i given X_i is negative binomial.
The regression coefficients θ_1, θ_2, , θ_p are approximated unknown parameters based on a collection of data.
Lawless [9] and L. Simon [14] provide a detailed explanation of the maximum likelihood estimation of the negative binomial regression model and the computation of related statistics. The moment estimation, which Breslow [18] initially proposed, is frequently used to calculate the parameters of the negative binomial model.
The importance of the null values and the heterogeneity of the corresponding population doesn't exclude the limits of the negative binomial model. For this reason, we attempted to apply the mixture regression model.

2.3 Proposed mixture regression model

The major problem in regressions is the choice of counting model, especially when we have a very heterogeneous data set. So the question that arises here: What is the most adequate and accessible regression model between the Poisson model and the Negative Binomial model? We adopt the "$K-means$" classification method to have an optimal selection. It is also well acknowledged that accident count or frequency statistics in the vehicle insurance industry frequently display zero mass and overdispersion, or extra-Poisson variation, which occurs when the response variable's variance exceeds the mean. When the Poisson model is applied incorrectly, the standard errors are underestimated and the significance of the regression parameters is overestimated, resulting in a misleading inference about the regression parameters. However, depending on the property of our dataset we obtained two classes using "$K-means$".(See Figure 1)
A first class that follows the Poisson distribution since we have in a situation of equidispersion between the mean and the variance of the response variable.
And a second class which follows the Negative Binomial distribution where we have an access of zeros and an overdispersion. This classification leads us to propose a "Poisson-NB" mixture regression model which gave us an accessibility to predict an optimal number of accidents for each new insured.
A crucial step in creating new probability distributions that can be utilized as a more adaptable substitute for conventional statistical distributions, particularly in overdispersion, is the mixing of probability distributions. We assume that the λ undergo a specific classification for the insured of an automobile insurance company, and that for a given individual, the distribution of accidents number Y follows a mixture of Poisson and Negative Binomial distributions.

Then, the probability mass function of Y that a randomly selected individual is given by:

$$\mathbb{P}(Y=y) \quad = \pi_1 Pois(y|\lambda_i) + \pi_2 NB(y|\mu_i, \alpha_i) \tag{4}$$

where:

.π_1: denotes the positive mixing weight of belongs to the Poisson class ($C0$).
.π_2=$1-\pi_1$: denotes the positive mixing weight of belongs to the BN class ($C1$).

2.4 Learning parameters of mixture model

We adopt the "$K-means$" classification method to have an optimal selection. Based on this classification, we obtain two classes ($K=2$). This last one, leads us to estimate the parameters of the proposed model, we use the EM algorithm. Given the cluster c_k denotes the k^{th} class, $y_i \sim f_k(.|\theta_k)$, then the mixture mass function of y_i is given by

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y|\theta_k).$$

The mixture model is a heterogeneous data model. It has a complex distribution of an observed variable $Y=(y_1, y_2, ..., y_n)$ given by (4).
Maximize the incomplete likelihood function:

$$\begin{aligned} l(y_1, y_2, ..., y_n|\Theta) &= \prod_{i=1}^{n}(\pi Pois(y_i|\lambda_{ix}) + (1-\pi)NB(y_i|\mu_{ix}, \alpha_{ix})) \\ &= \prod_{i=1}^{n}(\pi \frac{\exp(-\lambda_{ix})\lambda_{ix}^{y_i}}{y_i} \\ &\quad +(1-\pi)\frac{\Gamma(y_i+\alpha_{ix}^{-1})}{\Gamma(y_i+1)\Gamma(\alpha_{ix}^{-1})}(\frac{1}{1+\alpha_{ix}\mu_{ix}})^{\alpha_{ix}^{-1}}\alpha_{ix}^{-1}(\frac{\alpha_{ix}\mu_{ix}}{1+\alpha_{ix}\mu_{ix}})^{y_i}) \end{aligned}$$

with respect to $\Theta=(\alpha, \lambda, \mu, \pi)$ is a difficult task.
In order to estimate the mixture model parameters, we apply the EM algorithm. It is useful in a variety of heterogeneous or incomplete data problems. It is introduced by A.Dempster and al.[2], G.Maclachlan and al.[9]. The EM algorithm derives its name from the fact that each iteration operates two distinct steps: Expectation and Maximization steps.

1-Expectation step: The posterior probability such that the i^{th} observation y_i belongs to the k^{th} class (k=1,2), is calculated in the l^{th} iteration denoted by:

$$\begin{aligned} \tau_{ik}^{(l)} &= \mathbb{E}(z_{ik}|\Theta^{(l)}, Y_1=y_1, Y_2=y_2.....Y_n=y_n) \\ &= \mathbb{E}(z_{ik}|\Theta^{(l)}, Y_i=y_i) \\ &= \frac{\pi_k^{(l)} f_k(y_i|\Theta_k^{(l)})}{\sum_{k=1}^{K=2}\pi_k^{(l)} f_k(y_i|\Theta_k^{(l)})} \end{aligned}$$

with $i=1, ..., n$; $k=1,2$ and $\Theta^{(l)}=\Theta^{(l)}(y_1, ..., y_n)$ is the estimated parameter vector at the l^{th} iteration. Where,$f_k(y|\theta_k)$ denotes the mass function given the k^{th} cluster. For each observation y_i, we define the random vector

$Z_i = (z_{i1}, z_{i2}, ..., z_{ik})$ such that $z_{ik} = 1$ if y_i belongs to the class c_k. The random variables $(Z_i)_{1<i<n}$ are called latent variables or missing values which indicate the label class of y_i. Such that, $Z_i = (z_{i1}; z_{i2})$ denotes a dataset associated with $(y_1, y_2, ..., y_n)$. If $Z_{ik} = 1$ this means that the observation y_i belongs to the class $(k = 1, 2)$.

2-Maximization step: The conditional expectation of the complete data log likelihood $\mathbb{Q}$ is maximized with respect to the parameter vector denoted by

$$\begin{aligned}\mathbb{Q}(\Theta||\Theta^{(l)}) &= \mathbb{E}(logl_c(y_1, ...y_n, z_1, ...z_n|\Theta)|y_1, ...y_n, \Theta^{(l)}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K=2}\mathbb{E}(z_{ik}|\Theta^{(l)}; y_i)(log(\pi_k) + log(f_k(y_i|\theta_k)) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K=2}\tau_{ik}^{(l)}[log\pi_k + logf_k(y_i|\Theta_k)] \qquad (5)\end{aligned}$$

where $\Theta = (\pi_1, \pi_2; \Theta_1, \Theta_2)$ is the unknown parameters vector, $\Theta^{(l)} = (\pi_1^{(l)}, \pi_2^{(l)}; a_1^{(l)}, a_2^{(l)}; \Theta_1^{(l)}, \Theta_2^{(l)})$ is the parameters vector in the l^{th} iteration and l_c is the maximum likelihood function from complete data given by

$$l_c(y_1, ...y_n; Z_1, ...Z_n|\Theta) = \prod_{i=1}^{n}\prod_{k=1}^{K=2}\pi_k^{z_{ik}}f^{z_{ik}}(y_i|\Theta_k)$$

and the log-likelihood function from complete data is the follows,

$$logl_c(y_1, ...y_n; Z_1, ...Z_n) = \sum_{i=1}^{n}\sum_{k=1}^{K=2}[z_{ik}log(\pi_k) + z_{ik}log(f_k(y_i|\Theta_k)] \qquad (6)$$

According to equation (5), $Q(\mathbf{\Theta} \mid \mathbf{\Theta}^{(l)})$ is the expected value of the log likelihood function of $\mathbf{\Theta}$, with respect to the current conditional distribution of $\mathbf{z_1}, .., \mathbf{z_n}$ given $y_1, ..., y_n$ and the current estimates of the parameters $\mathbf{\Theta}^{(l)}$.

$$\mathbb{Q}(\Theta||\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K=2}\tau_{ik}^{(l)}[log\pi_k + logf_k(y_i|\Theta_k)] \qquad (7)$$

At the $(l+1)^{th}$ iteration, we have

$$\Theta^{(l+1)} = \underset{\Theta}{Argmax}\mathbb{Q}(\Theta||\Theta^{(l)}) \qquad (8)$$

In particular, the description of the EM algorithm [2] is given below where the parameters of the Poisson-NB mixture model are μ, α, λ and the mixing weight π.

Algorithm 1 The EM algorithm for Poisson-NB mixture model

1)Initialization: $\Theta^{(0)} = (\alpha^{(0)}, \lambda^{(0)}, \mu^{(0)}, \pi^{(0)})$

2) In the iteration l:

2-1 Expectation step:

$$\tau_{ik}^{(l)} \leftarrow \frac{\pi^{(l)} f_k(y_i|\Theta^{(l)})}{\pi^{(l)} f_k(y_i|\Theta^{(l)}) + (1-\pi^{(l)}) f_k(y_i|\Theta^{(l)})}$$

2-2 Maximization step:

$$\frac{\partial \mathbb{Q}}{\partial \mu} = 0 \leftarrow \mu^{(l+1)}$$

$$\frac{\partial \mathbb{Q}}{\partial \lambda} = 0 \leftarrow \lambda^{(l+1)}$$

$$\frac{\partial \mathbb{Q}}{\partial \alpha} = 0 \leftarrow \alpha^{(l+1)}$$

$$\frac{\partial \mathbb{Q}}{\partial \pi} = 0 \leftarrow \pi^{(l+1)}$$

3) If $||\Theta^{(l+1)} - \Theta^{(l)})|| < \epsilon$ is not satisfied return to step 2). where $\epsilon > 0$ is the threshold.

3 Actual Application

3.1 Data set

Let the variable y_i be the number of accidents of an individual i in a given period. Suppose that the number of accidents is independent from one individual to another. The set of these variables follow a Poisson law. The policyholder file contains a number of characteristics that can be used to explain and forecast the frequency, nature, and severity of auto accidents. These variables include details about the vehicle, including kind, power, the number of seats, and age, as well as individual-level information, like age, prior driving experience (Bonus-malus), driving license seniority, the number of drivers, and their location. These characteristics are outlined in the Table 1 below:

Covariates	Descriptions
X_1	Driver age
X_2	Vehicle age
X_3	Vehicle power
X_4	Bonus-Malus
X_5	Number of driver
X_6	Driving license seniority
X_7	Number of vehicle seats
X_8	Area

Table 1: Description of covariates

This example is based on data that we received from an insurer who runs an automobile insurance company in Tunisia. The main goal of this study resides in investigating the effect of a set of covariates (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8), on the number N of accidents counts. The data consists of 4908 insureds from the year 2019. The descriptive statistical analysis of all the explanatory variables in this study is shown in Table 2. The following (dependent variable) is the variable we're attempting to explain: Annual number of accidents involving a person's fault. It's a discrete variable with non-negative values that don't usually go beyond three accidents.(See Table 3)

	Mean	Std.Dev	Min	Max
N	1.446	1.343	0	3
X_1	4.178201	0.333421	3.232846	5.257965
X_2	1.320309	0.274465	0.682837	2.033251
X_3	5.634409	0.530544	4.243349	7.053131
X_4	3.911844	1.462189	1.259547	5.762619
X_5	2.642031	0.532639	1.278380	3.939587
X_6	1.366549	0.516778	0.524627	2.668481
X_7	7.269772	0.984514	5.215808	8.727790
X_8	2.636905	0.530625	1.169238	3.685506

Table 2: Table illustrating the descriptive statistics of data

Number of claims	Frequency	percentage	cumulative frequency	cumulative percent
0	0.505349	50.5349	0.870352	87.0352
1	0.365002	36.5002	0.365002	36.5002
2	0.129648	12.9648	0.999999	99.999
3	0.000321	00.0321	1.00032	100.032

Table 3: Proportional frequency of claims

3.2 Results and discussion

Using data from the files of our private insurer, we will first try to answer the following question: what is the statistical relationship between the number of accidents of an individual, his personal qualities and the characteristics of his vehicle? And secondly, what is the most appropriate and robust model for fitting the data set?

Models	Poisson		NB	
Covariables	Coefficient	P-Value	Coefficient	P-Value
Intercept	−5.2779	0.000	−6.3149	0.000
$X1$	−0.6581	0.000	−0.6305	0.000
$X2$	−0.4256	0.000	−0.3620	0.034
$X3$	0.1175	0.097	0.1436	0.173
$X4$	1.0256	0.000	1.0598	0.000
$X5$	−0.1144	0.123	−0.0822	0.453
$X6$	−0.1542	0.166	−0.1792	0.271
$X7$	0.3112	0.000	0.3595	0.000
$X8$	0.2848	0.000	0.3034	0.005

Table 4: Coefficients regression and their P-Values for Poisson and NB models

Both the GLM Poisson and GLM negative binomial regressions indicate the same explanatory variables for claim frequency, with similar effects. Based on the P-Value analysis, the two models (GLM Poisson regression and GLM NB regression) indicate that five main effects (Driver age X1, Vehicle age X2, Bonus-malus X4, Number of vehicle seasts X7, Area X8) are significant (P-Value<0.05), but regarding the best significance for (X2, X8) , the GLM Poisson regression model shows its effectiveness as shown in the Table (4). We notice an increase of the claims with the Bonus-Malus coefficient ($X4$), which is natural since it reflects the past of the driver, moreover the coefficient number of vehicle places ($X7$) also it leads to an increase, since the latter differs according to the category of use, i.e. use of a family car (business of 5 places) is not the same as utility use (commercial of 2

places) and use of transport of persons (cab =4 places, renting = 9 places and rural transport =8 places). Table (4) shows that there is a significant correlation and a positive effect between the area factor ($X8$) and the response variable (number of claims 'N'), which translates into the importance of road conditions in reducing the number of accidents. In this study, we compare the infrastructure of the northern and southern countries of Tunisia and we find that the more we have a good road structure, the more the drivers are comfortable in driving and consequently less road accidents.

It is now known which factors have a significant effect on the number of accidents. But what we are interested in here is how to rank a new insured from its own coordinates to estimate its number of accidents in order to predict its optimal premium. To do this, we rely on the $K-means$ method as indicated in the previous section and on GLM Poisson and NB regression to predict an optimal number of accidents for each insured. From the criteria of choice of model performance, Table (5) shows that GLM Poisson regression is the best model compared to GLM NB regression for the first class (C0). On the other hand, GLM NB regression shows its performance for the second class (C1).

Class	C0		C1	
Models	Poisson	NB	Poisson	NB
$Deviance$	0.51237	7.4449	161.48	70.191
AIC	32.5124	51.4417	82.8	67.6
BIC	81.89191	106.31	243.269	152.02
$Pearson-chi2$	0.281	7.47	165	73.1

Table 5: Criteria for each class's comparison with Poisson and NB regression models

Criteria	Poisson	NB	Poisson-NB Mixture
MSE	0.05817	0.05261	0.0153
AIC	676.6	956.4	432.11
BIC	735.09	821.39	441.96
$Errorrate$	0.044	0.037	0.0007

Table 6: Model selection criteria

The MSE, AIC, BIC, and error rate values utilized in model selection are shown in Table(6). The model with the smallest error value is the best fit. In comparison to the other distributions discussed here, the Poisson-NB regression mixture distribution is the best at fitting the data set, as shown by the above results (GLM

Poisson regression, GLM NB regression). The results showed that in the classification problem, the mixture distribution is a formidable contender.

4 Conclusions and perspectives

The key goal for all auto insurance firms is to forecast the number of accidents for each new insured in order to have an accurate classification. The important role of this paper is to use the mathematical tools and learning methods necessary to achieve this goal. Among the best known classification methods, we used the $K-means$ method to have the most possible heterogeneous classes between them, through an inter classification and an intra classification. In this paper, we used the $K-means$ as a machine learning method to classify our data sets, as well as a regression method to estimate explanatory variables and predict the number of accidents. Using machine learning approaches, we can divide insureds into two groups (K=2) based on the number of incidents they have had and the risk models they have used. For the insurer, this methodology identifies a two-class solution:
-First class is the most dangerous; it is reserved for insureds who have had more than one accident during the exercise period.
-The second class is the least dangerous; it is for insureds who have never had an accident or have had only one accident during the exercise period.

The results show that when compared to the other regressions, the Poisson-NB mixture regression has a lower AIC; error rate and minimal MSE value. This demonstrates that the proposed mixture model is the best fit for estimating policyholder accidents.

The acquired results can help to increase accuracy of the extended mixture regression model used to estimate the coefficients of the level of premiums in the Tunisian Bonus-malus table, which can be used to predict the premium paid by the insured.

Bibliography

[1] A. E. Renshaw, *Modelling the Claims Process in the Presence of Covariates*, ASTIN Bulletin, Vol. 24, No. 2, 265-285, (1994).

[2] A. Dempster, N. Laird, and D. Rubin, *Maximum likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society B (Methodological), Vol. 39, No. 1, pages 1-38, (1997).

[3] C. Akantziliotou, R. A. Rigby, and D. M. Stasinopoulos, *A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution*, Comput. Stat. Data Anal., 53, 381-393, (2008).

[4] S. Aryuyuen and W. Bodhisuwan, *The negative binomial-generalized exponential (NB-GE) distribution*, Appl. Math. Sci., 7, 1093-1105, (2013).

[5] J. Del Castillo and M. Pérez-Casany, *Overdispersed and underdispersed Poisson generalizations*, J. Stat. Plan. Inference, 134, 486-500, (2005).

[6] T. C. Fung, A. L. Badescu, and X. S. Lin, *A class of mixture of experts models for general insurance: Application to correlated claim frequencies*, ASTIN Bulletin: The Journal of the IAA, 49(3), 647-688, (2019a).

[7] T. C. FUNG, A. L. BADESCU, AND X. S. LIN, *A class of mixture of experts models for general insurance: Theoretical developments*, Insurance: Mathematics and Economics, 89, 111-127, (2019b).

[8] W. H. GREENE, *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*, Working Paper EC-94-10, Department of Economics, Stern School of Business, New York University, (1994).

[9] G. MACLACHLAN AND T. KRISHNAN, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, 2nd edition, (2007).

[10] J. GARRIDO, C. GENEST, AND J. SCHULZ, *Generalized linear models for dependent frequency and severity of insurance claims*, Insurance: Mathematics and Economics, 70, 205-215, (2016).

[11] J. HINDE AND C. G. B. DEMÉTRIO, *Overdispersion: Models and Estimation*, Associacao Brasileira de Estatistica, Sao Paulo, (1998).

[12] H. ZAMANI AND N. ISMAIL, *Negative Binomial-Lindley Distribution And Its Application*, J. Mathematics And Statistics, 1, 49, (2010).

[13] J. F. LAWLESS, *Negative binomial and mixed Poisson regression*, The Canadian Journal of Statistics, Vol. 15, No. 3, Pages 209-225, (1987).

[14] D. LUKASZ, L. MATHIAS, AND M. V. WÜTHRICH, *Gamma Mixture Density Networks and their application to modelling insurance claim amounts*, Insurance: Mathematics and Economics, Vol. 101, Part B, Pages 240-261, (2011).

[15] L. SIMON, *Fitting negative binomial distribution by the method of maximum likelihood*, J. Casualty Actuarial Society, 17, 45-53, (1961).

[16] L. BREIMAN, J. FRIEDMAN, R. OLSHEN, AND C. STONE, *Classification and regression trees*, Wadsworth Brooks, (1984).

[17] B. G. LINDSAY, *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics 5, Hayward: Institute of Mathematical Statistics, (1995).

[18] L. BREIMAN, *Random forests*, Machine Learning, 45, 5-32, (2001).

[19] M. AITKIN, D. ANDERSON, B. FRANCIS, AND J. HINDE, *Statistical Modelling in GLIM*, Oxford University Press, New York, (1990).

[20] G. MCLACHLAN AND D. PEEL, *Finite Mixture Models*, Wiley Series in Probability and Statistics, John Wiley and Sons Inc., (2000).

[21] M. LICHMAN, *UCI Machine Learning Repository*, http://archive.ics.uci.edu/ml, (2013).

[22] N. E. BRESLOW, *Extra-Poisson Variation in Log-Linear Models*, Journal of the Royal Statistical Society Series C, Royal Statistical Society, vol. 33(1), pages 38-44, (1984).

[23] R. VERBELEN, L. GONG, K. ANTONIO, A. BADESCU, AND S. LIN, *Fitting Mixtures Of Erlangs To Censored And Truncated Data Using The EM Algorithm*, ASTIN Bulletin: The Journal of the IAA, Vol. 45, Issue 3, pp. 729-758, (2015).

[24] S. M. GOLDFELD AND R. E. QUANDT, *A Markov model for switching regressions*, Journal of Econometrics, Vol 1, Issue 1, Pages 3-15, (1973).

[25] S. C. K. LEE AND X. SHELDON LIN, *Modeling Dependent Risks with Multivariate Erlang Mixtures*, ASTIN Bulletin: The Journal of the IAA, Volume 42, Issue 1, pp. 153-180, (2012).

[26] P. SHI AND E. A. VALDEZ, *Multivariate negative binomial models for insurance claim counts*, Insur Math Econ 55, 1829, (2014).

[27] K. F. SELLERS AND A. RAIM, *A flexible zero-inflated model to address data dispersion*, Comput. Stat. Data Anal., 99, 68-80, (2016).

[28] S. C. TSEUNG, A. BADESCU, T. C. FUNG, AND X. S. LIN, *LRMoE.jl: a software package for insurance loss modelling using mixture of experts regression model*, Ann. Actuar. Sci., 15(2), 419-440, (2021).

[29] W. JIANG AND M. A. TANNER, *On the Approximation Rate of Hierarchical Mixtures-of-Experts for Generalized Linear Models*, Neural Computation, Vol 11, Issue 5, 1183-1198, (1999).

[30] R. WINKELMANN, *Econometric Analysis of Count Data*, Springer-Verlag, (2003).

[31] K. K. Yau, K. Wang, and A. H. Lee, *Zero-Inflated Negative Binomial Mixed Regression Modelling of Over-Dispersed Count Data with Extra Zeros*, Biometrical Journal, 45, 437-452, (2003).

[32] Z. Yang, J. W. Hardin, C. L. Addy, and Q. H. Vuong, *Testing Approaches for Overdispersion in Poisson Regression versus the Generalized Poisson Model*, Biometrical Journal, 49, 565-584, (2007).

[33] Y. Lv, S. Tang, and H. Zhao, *Real-time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method*, International Conference on Measuring Technology and Mechatronics Automation, pp. 547-550, (2009).

[34] H. Zamani, P. Faroughi, and N. Ismail, *Bivariate generalized Poisson regression model: applications on health care data*, Empir Econ, 51(4), 1607-1621, (2016).

[35] P. Zhang, D. Pitt, and X. Wu, *A new multivariate zero-inflated hurdle model with applications in automobile insurance*, ASTIN Bulletin: The Journal of the IAA, 124, (2022).

How to Cite: Maryem Jaziri[1], Afif Masmoudi[2], *Modeling auto insurance frequency using K-means and mixture regression*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 3, No. 2, Pages:95–111, (2023).