JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ESTIMATION OF NUMBER OF CLAIMS IN MOTOR INSURANCE USING GENERALISED LINEAR MODELS and ARTIFICIAL NEURAL NETWORKS.

[1]Ramsai Jagadish Yenugadhati, [2]Bala Nagendra Babu, [3]Satya Sai Baba Mudigonda, [4]Rohan Yashraj Gupta, [5]Phani Krishna Kandala,
[6]Pallav Kumar Baruah
[1]M.Sc Data Science & Computing, [2]M.Sc Data Science & Computing, [3]Adjunct Professor, [4]Visiting Faculty, [5]Visiting Faculty, [6]Professor,
Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prashanti Nilayam, India

Abstract : The profitability and competitiveness of the motor insurance business are directly dependent on accurately estimating the claim count. Statistically, it can be seen as a Regression problem. Claim count is a discrete random variable that allows only discrete positive values which is count data. Negative Binomial (NB2), Poisson, Quasi-Poisson, and Artificial Neural Network Models are used for Regression analysis on discrete data. For this work crash analysis system (CAS) dataset is used. The performance of the models is measured using various Error metrics. It is found that the Negative Binomial regression model performs better than the other models with a mean square error of 0.09 and a root mean square error of 0.30.

Index Terms –
Estimation, Generalized Linear Model, Negative Binomial Distribution, Poisson Distribution, Quasi-Poisson Distribution, Artificial Neural Network, Distributions, Count data, Regression, Error metrics.

I. INTRODUCTION

In this work, we study the estimation of claim counts for Motor Insurance. There is room for reform in the insurance industry because of the notion of forecasting risk and uncertainty. The price of insurance contracts is determined using mathematical, statistical, and domain-related methodologies. The fundamental purpose is to help policyholders with claims while being profitable. To achieve this, the company must be accurate with the Estimation in terms of the number of claims. In a competitive market, an insurance firm that does not provide premiums based on individual risk factors would not survive.
As depicted in the figure 1.1 The estimation of the number of claims for the Motor Insurance business is addressed earlier using various platforms and various packages.

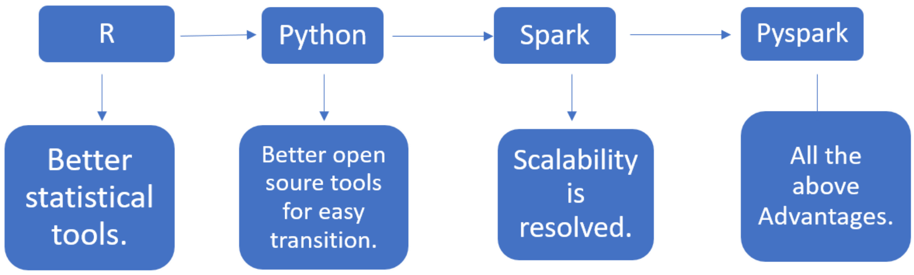

Fig: 1.1 Different Platforms Implementation

In this work, Crash Analysis System (CAS) dataset is used. The flow of this work can be illustrated using the figure 1.2. At first, the missing values handling is done which comes under data preprocessing. Spark's data preprocessing pipeline is used for the steps like indexing, Encoding, and Scaling. The Recursive feature Elimination process is an automated feature selection process for selecting the list of important features from the total number of available features. Later, the resultant dataset is passed on to the modelling phase involving negative binomial, Poisson, Quasi-Poisson regression, and neural network implementations. The best fit model is chosen comparing the regression metrics root mean square error and mean absolute error.

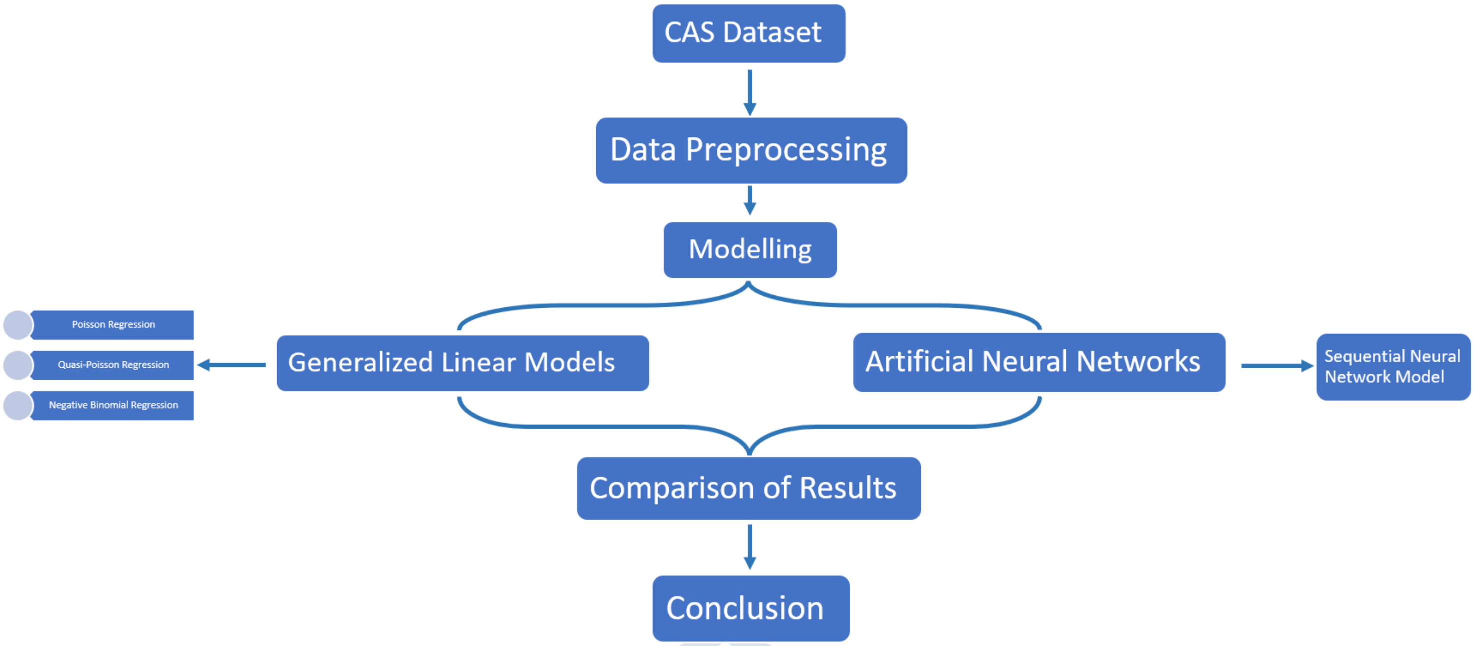

Fig: 1.2 Flow of the Work

II. LITERATURE SURVEY

Researchers have offered many works and approaches in the area of insurance and generalized linear models on various platforms. This study heavily relies on generalized linear models. The authors Raghavendra KS and Dr. Bheemanagouda explored the relevance of actuarial science and actuaries in the insurance business in their research paper "Study on Actuarial Science and Actuaries-The Pioneer of Insurance. [1]" The research features and functions of this topic are discussed. The process of developing insurance products and aspects to consider, as well as the implications for pricing strategies, were discussed. Achim Zeileis, Christian Kleiber, and Simon Jackman describe different extended linear models and their implementation in the R language in their research article "Regression Models for Count Data in R. [2]" The models are applied to "demand for medical care in the elderly," a dataset created for an R package. The hurdles and zero-inflated models are implemented using the MASS package from CRAN. The authors examined the use of GLMs to fit data in cases of overdispersion in their work “Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial[16], and zero-inflated generalized Poisson regression models. [3]" The authors utilized the datasets "German Socio-economic panel data (GSOEP)" and "Malaysian own damage claim counts" as references in this work. Amy Byers, Thomas Gill, and Heather G Allore explored the better fit of the Negative binomial model on discrete data with the problem of overdispersion in their research work " Application of Negative Binomial Modelling for Discrete Outcomes.[4]" The information comes from a clinical experiment that was created to "assess the effectiveness of a rehabilitation program." Authors Ashenafi A. Yirga1, Sileshi F. Melesse, Henry G. Mwambi, and Dawit G. Ayele examined the use of GLMs in the biomedical area in their work "Negative binomial mixed models for assessing longitudinal CD4 cells count data.[5]" For fitting overdispersed data, Poisson regression and negative binomial regression models are utilized, as well as hurdle and zero-inflated models for data with excess zeroes. The data used in this investigation came from CAPRISA 002 AI.

The authors of the academic work "Case Study: French Motor Third-Party Liability Claims,[6]" Alexander Noll, Robert Salzmann, and Mario v. Wuthrich integrated claim count and claim frequency modelling utilizing Machine learning and Data Science methodologies. "Predictions of that third party liability insurance dataset demonstrate superior accuracy utilizing machine learning and neural network technique as compared to generalized linear models,[7]" the researchers concluded. BMC Health Services Research Ahmed Nabil Shaaban, Bárbara Peleteiro, and Maria Rosario O. Martins Several strategies were used to select the best count fit model in the work "[8]Statistical models for analyzing count data: predictors of length of stay among HIV patients in Portugal using a multilevel model," including the Poisson regression model, zero-inflated Poisson, negative binomial regression model, and zero-inflated negative binomial regression model. The duration of stay among HIV hospitalizations in Portugal was predicted using a total of 26,505 individuals classified under the Major Diagnostic Category (MDC) developed for patients with HIV infection, with HIV/AIDS as a primary or secondary cause of admission. Between January 2009 and December 2017, the registered discharges at Portuguese National Health Service (NHS) institutions. Modelling claim numbers when there are more zeros in the data, Claim frequency statistics in general insurance may not always follow the typical Poisson distribution, and they may be zero-inflated in some cases. Under the Poisson or even the negative binomial distribution assumptions, extra dispersion is defined as the number of observed zeros surpassing the number of predicted zeros. As goodness-of-fit and model selection metrics, the generalized Pearson chi-square statistic, Akaike's information criteria (AIC), and Bayesian information criteria (BIC) are utilized. Claim frequency and claim size in non-life insurance using spatial modelling, This paper discusses non-life insurance claim frequency and average claim size models. The objective of GoncaBuyrukoglu, SelimBuyrukoglu, and Zeynal Topalcengiz's work "[29]Comparing Regression Models with Count Data to Artificial Neural Network and Ensemble Models for Prediction of Generic Escherichia coli Population in Agricultural Ponds Based on Weather Station Measurements" was to see how well zero-inflated Poisson and hurdle negative binomial regression models with count data compared to artificial neural networks and ensemble models The mean absolute error (MAE) value across six pre-tested models was used to decide the use of artificial neural networks, AdaBoost, and random forest. The MAE was also utilized to evaluate two-part count data models for artificial neural network and ensemble models.

Various insights are taken from the research work which was done earlier by many researchers in handling count data using various generalized Linear models and Artificial Neural Networks.

III. METHODOLOGY

3.1 Data Description and EDA

This information comes from the Waka Kotahi Crash Analysis System[9] (CAS), which keeps track of all road accidents reported by the New Zealand Police, which is the publicly available dataset. CAS dataset covers accidents on all highways of New Zealand and sites where the general public has unrestricted access to operate a vehicle. This dataset consists of 7,58,757 rows and 71 columns. Figure 3.1.1 depicts Different levels of crash severities are mentioned in the Crash Analysis System dataset. They are “Non-Injury,” “Minor Injury,” “Serious Injury,” and “Fatal Injury.” From the figure 3.1.1 we can see that the non-injury category is high in number.

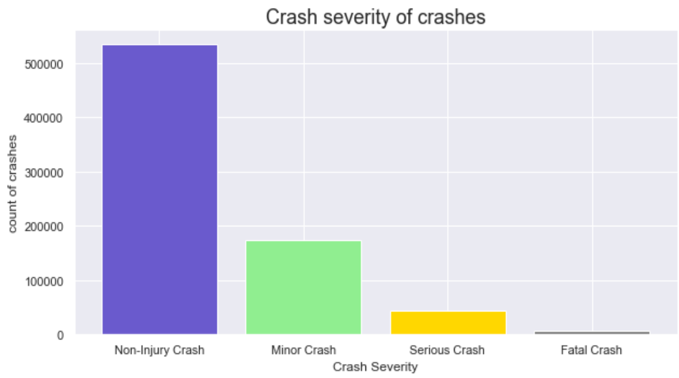

Fig: 3.1.1 Levels of Crash Severity

Table 3.1.1 is the data dictionary of few important features from the Crash Analysis System Dataset.

Feature	Description.
Speed	The number of accidents or crashes is directly proportional when it comes to speed and speed limits. We have two columns, AdvisorySpeed and Speed Limit. Both tell about what is the advisory speed at the point of the crash and what the limit is at that point of crash.
Vehicle types	A variety of vehicle types are reported in the crash, and most of the time, the number of accidents that take place is zero. When the vehicle type is unknown, it is labeled “vehicle.” 'bicycle,' 'bus,' 'car Station Wagon', 'moped', 'motorcycle', 'other Vehicle Type', 'SUV', 'schoolBus', 'taxi', 'train', 'truck', 'van or utility, 'vehicle' are the columns.
Weather conditions	The number of crashes can be related to weather conditions as well. The crash analysis system dataset has two columns stating what the weather condition was at that particular moment of the crash. We also see in this particular dataset harsh climate conditions did not affect much, or the number of crashes is less even in bad weather conditions.
Regions	Accidents are more in number in a few places when compared to other places. Indicating the name of the region, we have two columns in this crash analysis system dataset stating which place is that and what its id is. Auckland region has the highest number of reported crashes.
Road character	The general character of the road. Among the possible values are 'Bridge,' 'Motorway Ramp,' 'Rail crossing,' and 'Nil.'
Road surface	The road surface description used at the crash.' Unsealed' or 'sealed' are possible values.
road curvature	The road's curvature has been simplified. 'Curved' and 'Straight' are two possible values.
Urban	A variable is derived from the 'spdlim' variable. 'Urban' (urban, spd lim 80) or 'Open Road' (open road, spd lim >=80 or 'LSZ') are possible values.
Traffic sign	The number of times 'traffic signage' (including traffic poles, signals, roadside delineators, or bollards) was struck in the crash.
Car station wagon	As a variable, the number of automobiles or station wagons involved in the collision was calculated.
Cliffbank	The number of times a 'cliff' or 'bank' was impacted during the crash. That is a derived variable.
Debris	This explains the number of times in the accident that rubble, stones, or things dropped or flung from vehicles were impacted.
Ditch	This derived variable indicates the number of times a 'ditch' or 'waterable drainage channel' was impacted in an accident.

Fence	This derived variable indicates the number of times a 'fence' was hit during the accident. (other similar things that fall within this category are Letterboxes, hoardings, private roadside furniture, hedges, and sight rails.)
Guard rail	This derived variable denotes a crash. 'New Jersey' barriers, 'ARMCO' barriers, sand-filled barriers, wire-catch fences, and other similar devices fall into this category.
Motorcycle	The number of motorcycles involved in the crash was derived.
Number of lanes	The number of lanes on the crash road (number).
Post or pole	The number of times a post or pole was struck during the crash is indicated by this derived variable. This includes Light, power, phone, and utility poles, as well as objects that are practically part of a pole.
School bus	The number of buses involved in the accident, except school buses, which are listed in the 'schoolbus' column. That is a derived variable.
Speed limit	The speed (spd) and limit were in effect at the time of the crash. It could be a number or the abbreviation 'LSZ' for a Limited speed zone.
Tree	The number of times trees or other growing items, were struck during the crash is indicated by this derived variable.
Vehicle	The number of times a stationary attended vehicle was struck in the crash was derived as a variable. This includes broken down vehicles, workers' vehicles, taxis, and buses.

Table 3.1.1 Data Description

The table 3.1.2 gives the values of number of claims (which correspond to the number of serious injuries in a crash) versus the number of incidents taken place throughout the New Zealand. Excluding the number of zero claim counts present in our target variable(that case being the highest), there is a high imbalance in the dataset.

1	40180
2	4334
3	863
4	258
5	81
6	26
7	8
8	5
9	1
10	3
12	1
14	1

Table 3.1.2 Number of serious injuries in a crash vs Number of Incidents

The figure 3.1.2 is the pictorial representation of the Serious injury crashes distributed to the number of incidents taken place excluding the case of no serious injuries in a incident.

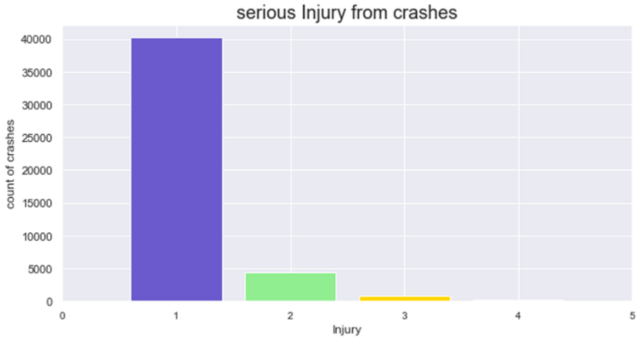

Fig: 3.1.2 Class imbalance in target variable

3.2 Data Preprocessing

Figure 3.2.1 is the illustration of spark pipeline developed for data preprocessing. Important conversions like string indexing, one hot encoding are executed as per requirement for the existing categorical features which are selected in the previous Recursive feature elimination phase.

a) Drop null value

For the columns in which missing values exceed more than or equal to a certain threshold, they are to be dropped as it is meaningless to create the artificial data points from nowhere. In this work, columns having more than 90% missing values are dropped.

b) Handle missing values with imputation.

Now, try imputing the values in the remaining places where it is feasible using various values such as mean, median or mode, or zeroes if it makes sense. Simple zero value imputation is used in this work as most of the time it is not likely that a pole or a tree is involved or some particular vehicle is involved in the crash.

c) Outlier detection, removal, and imputation

In the next step try detecting the outliers as they impact the model's performance. Handling them is also an important step. Outliers cannot be more in number so that they can be dropped, or maybe they can be limited to a certain threshold. Claims Count anything more than 10 are considered to be 10.

d) Drop feature/column

Few features may not be helpful for the model, and those columns can be dropped from the list of columns used in modelling. Selection of those columns can initially be made manually, and later, a few algorithms can be applied to select the columns based on criteria like numerical or categorical, etc. to drop the features that do not contribute much to the model in predicting the target variable are dropped using an automated feature selection process which is recursive feature elimination using Poisson regressor.

e) Convert categorical feature/column into numerical

One hot encoding and string indexing are two features that can be part of the data preprocessing pipeline of any spark modelling procedure. The columns with value being of string type cannot be used in algorithms, which are meant to be changed into the numerical format.
For generalization and automation, we separate the numerical and categorical labels into separate lists and then pass the features into the pipeline.

Fig: 3.2.1 Data preprocessing Pipeline

3.3 Automated Feature Selection:

Feature selection is an integral part of the machine learning model construction process. It is the process of automatically or manually selecting a subset of the most valuable and relevant features to be used in model creation. Feature engineering is separated into two steps: feature selection and feature extraction, which is required for any machine learning approach. Despite having the same purpose, feature selection and extraction are fundamentally different procedures. The main difference is that feature selection picks a subset of the original feature set, whereas feature extraction creates new ones. To avoid overfitting, feature selection is a strategy of restricting the model's input variable by only using relevant data.

Wrapper technique:

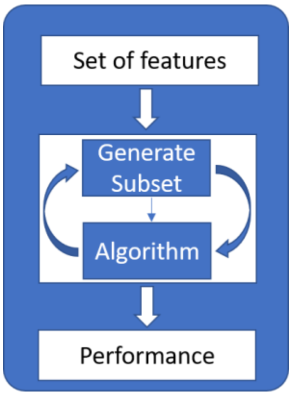

Fig 3.3.1 Wrapper model for Feature Selection

It is a feature selection problem where the search is issued in which many combinations are created, assessed, and compared to other combinations. It iteratively trains the algorithm by utilizing the subset of characteristics. Based on the model's output, features are added or removed, and the model is then retrained with this feature set.

Recursive-Feature Elimination:

It's a "Recursive greedy optimization method". As depicted in the figure 3.3.1, in which the features are chosen by taking a smaller and smaller subset of features iteratively[7][11]. With each set of features, an estimator is trained, and the significance of each feature is calculated using the coef_attribute or a feature_importances_attribute. We implemented the Recursive Feature Elimination Method to select significant features for our work. The number of selected features can be of any number, but we used the most significant top 10 features of the processed data for our model. In our work from an exhaustive list of 71 features, the selected features are 'roadCharacter[T.Motorway ramp]', 'carStationWagon', 'fence', 'guardRail', 'motorcycle', 'NumberOfLanes', 'parkedVehicle', 'postOrPole', 'speedLimit', 'tree'.

IV. GENERALISED LINEAR MODELS

GLMs have three components, which are comparable to but not identical to the components of a linear regression model.
In specific, GLMs are made using the following components:

- All observations for the target variable are assumed to be independent and selected from an exponential family distribution.
- X1, X2,..., Xk; vector of 'k' input independent variables and a vector of k+1 parameters.
- (b0, b1,..., bk) and a link function g(), which allows us to define g(E(Y)) as a linear combination of our input variables, with m = E.(Y).

The probability of the levels of a categorical response variable being transformed via a link function into an unbounded continuous scale. When the transformation is finished, linear regression may be used to model the connection between the predictors and the response. A binary response variable, for example, can only have two values. After converting those values to probabilities, the response variable now has a range of 0 to 1. The outcomes of applying an appropriate link function to such probabilities vary from to $-\infty$ to $+\infty$. Link function can be generalised to $g(\mu i) = Xi'\beta$. Any link function's objective is to change the output variable so that it may be expressed as a linear combination of the input variables. Certain link functions are typically applied depending on the probability distribution from which we expect our output distribution to be taken

In our analysis, we are using three statistical distributions which are apt for modelling the count variable 'seriousInjuryCount' in our dataset. The three distributions are

1. Poisson Distribution
2. Quasi-Poisson Distribution
3. Negative Binomial Distribution

4.1 Modelling using the Poisson Regression Model

"The coefficient estimate indicates the average change in the log odds of the response variable associated with a one-unit increase in each predictor variable"[12]. coefficients are all about how the dependent variable varies with respect to each independent variable. Here are the features: motorcycle, Light Overcast, and Roadlane 2-way have more effect on the dependent variable 'seriousInjuryCount' compared to all other independent variables.

	coef	std err	z	P>|z|	[0.025	0.975]
roadCharacter[T.Motorway ramp]	-0.9666	0.065	-14.833	0.000	-1.094	-0.839
carStationWagon	-0.7083	0.006	-111.322	0.000	-0.721	-0.696
fence	-0.1051	0.015	-6.899	0.000	-0.135	-0.075
guardRail	-0.0731	0.027	-2.683	0.007	-0.126	-0.020
motorcycle	0.7559	0.011	68.241	0.000	0.734	0.778
NumberOfLanes	-0.6785	0.006	-118.082	0.000	-0.690	-0.667
parkedVehicle	-0.9405	0.024	-39.849	0.000	-0.987	-0.894
postOrPole	-0.0582	0.019	-3.032	0.002	-0.096	-0.021
speedLimit	-0.0053	0.000	-34.161	0.000	-0.006	-0.005
tree	0.3205	0.018	18.089	0.000	0.286	0.355

Figure 4.1: Poisson Model

The figure 4.1 is the output of the Poisson regression model "The standard error gives us an idea of the variability associated with the coefficient estimate. We then divide the coefficient estimate by the standard error to obtain a z value"[12]. In our model, we have all the standard error values to be around 0.01 and it is not very large which implies that the coefficient estimate is near to precision. "The p-value Pr(>|z|) tells us the probability associated with a particular z value. This essentially tells us how well each predictor variable can predict the value of the response variable in the model"[12].We can determine the significance level based on our preferences be it 1%, 5%, or 10%. As a 0.05 level of significance is common across models implemented using the Generalized linear models, we have all the p-values less than this threshold of 0.05, we take it that the independent variables used in this model are quite significant and are contributing well to the regression model. "The null deviance in the output tells us how well the response variable can be predicted by a model with only an intercept term. The residual deviance tells us how well the response variable can be predicted by the specific model that we fit with p predictor variables. The lower the value, the better the model can predict the value of the response variable"[12]. If our data shows overdispersion, the use of a Poisson model will underestimate the standard errors which will result in too low p-values with an increased risk for a type I error. As the difference between estimated variance and estimated mean is considered as the dispersion, in this case, we consider dispersion to be 1 as the mean and the variance of the Poisson model assumes to be equal. Therefore, it makes the dispersion value be 1.

4.2 Modelling using Quasi-Poisson Regression Model

Figure 4.2.1 is the output of the Quasi Poisson model[22,23], Which will have similar coefficients but will have different standard errors and also better p-values. In the Poisson model, the model overlooks this dispersion value, as it considers it to be 1 irrespective of the original value since it upholds the assumption that the mean and variance are equal. Thereby the chances of type I error are maximum.

```
QuasiPoisson GLM Model Summary.
===================================================
Name                 Parameter Estimate  Standard Error
---------------------------------------------------
Intercept                         -3.82            0.02
regionC                            0.18            0.01
roadLane2                          0.61            0.01
streetLight0                      -0.27            0.01
trafficControlU                    0.05            0.01
bicycle                            0.97            0.01
carStationWagon                   -0.28            0.00
motorcycle                         0.98            0.01
NumberOfLanes                     -0.12            0.00
speedLimit                         0.02            0.00
tree                               0.50            0.01
```

Figure 4.2.1: The Quasi Poisson Model

As we can observe, the Coefficients of both the models' outputs are very similar to what we had discussed earlier. The varying points are only the standard errors and p-values. The Quasi-Poisson model will result in the same coefficients as the Poisson model, but with different standard errors and p-values since it adjusts for under or overdispersion. Once the model is obtained, the next part is to look at the predictions using the test data. As the model is already trained, we compare the actual values and the predicted values in the figure 4.2.2.

	Actual	Predicted	Residuals
549049	0.0	0.046926	-0.046926
444549	1.0	0.069334	0.930666
518123	0.0	0.053422	-0.053422
437517	0.0	0.069366	-0.069366
750131	0.0	0.024980	-0.024980
772252	0.0	0.560544	-0.560544
440714	3.0	0.254505	2.745495
25970	0.0	0.031563	-0.031563
313461	0.0	0.034000	-0.034000
531636	0.0	0.028879	-0.028879

Figure 4.2.2 Actual vs Predicted

Randomly, 10 rows are selected and these are the actual and predicted values of the Poisson model obtained before. Here, Predicted values are looking to be near zero in almost all the cases. Figure 4.2.3 is the residual plot for interpreting how residuals are distributed.

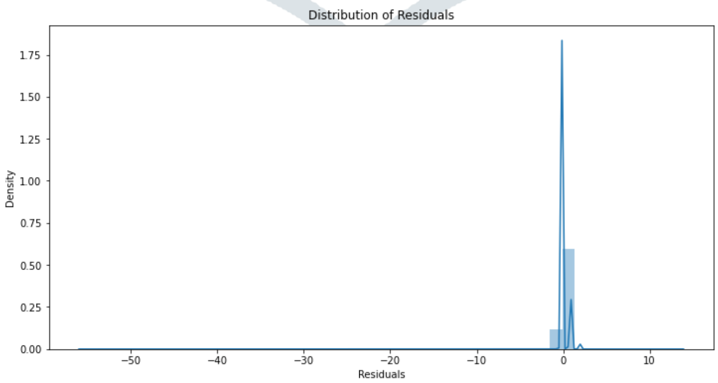

Figure 4.2.3 Residuals plot

4.3 Modelling using Negative Binomial Model

In this work, Negative Binomial Regression among Generalized Linear Models is used to build a regression model for predicting the number of claims. g(m) = ln(m) is the link function for negative binomial regression. After accounting for the influence of the predictors, one important criterion is the relative value of the variance to the mean. Overdispersion can be detected using the Pearson Chi2 dispersion statistic, residual plots, or a Likelihood Ratio Test. α known as the "overdispersion" parameter[30][31], is quite valuable. When it is near 0, the mean and variance are equal, indicating that our distribution is a Poisson distribution with a single parameter μ. The Pearson Chi2 dispersion statistic: the dispersion parameter φ can be estimated using Pearson's Chi-squared statistic and the degree of freedom. When φ is more significant than 1, it is overdispersion. Linear regression is generalized by allowing a link function to be used to link the linear model to the response variable. The below figure 4.3.1 depicts the presence of over dispersion in the model.

```
 Overdispersion test

data:  regp
z = 34.459, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.307874
```

Fig 4.3.1 Overdispersion test

Regression approach for NB2 in a nutshell:

1. On the data set, apply the Poisson regression model, and that will give us the fitted rate vector.
2. On the data set, fit the OLS regression model. This will provide us with the value of α.
3. Fit the NB2 regression model to the data set using the α from STEP 2.
4. Using the fitted NB2 model, make predictions about projected counts on the test data set.
5. Check the NB2 model's goodness of fit.

Below given figure 4.3.2 is the output of the final model among the GLMs. In this Approach, the lambda vector is taken from the trained Poisson regression model. Later that vector is used in estimating the appropriate "α" using implementing ordinary least squares regression. After verifying the parameter with the Student's T-test of 0.05 significance and degrees of freedom, the computed α is used to train the Negative Binomial Regression model.

	coef	std err	z	P>|z|	[0.025	0.975]
roadCharacter[T.Motorway ramp]	-0.9467	0.065	-14.557	0.000	-1.074	-0.819
carStationWagon	-0.7161	0.006	-112.451	0.000	-0.729	-0.704
fence	-0.1042	0.015	-6.828	0.000	-0.134	-0.074
guardRail	-0.0937	0.028	-3.387	0.001	-0.148	-0.039
motorcycle	0.7330	0.010	72.003	0.000	0.713	0.753
NumberOfLanes	-0.6723	0.006	-116.285	0.000	-0.684	-0.661
parkedVehicle	-0.9276	0.024	-39.219	0.000	-0.974	-0.881
postOrPole	-0.0666	0.019	-3.454	0.001	-0.104	-0.029
speedLimit	-0.0054	0.000	-34.767	0.000	-0.006	-0.005
tree	0.3285	0.018	18.542	0.000	0.294	0.363

Fig 4.3.2 Negative Binomial Model

4.4 Generalized Linear Models Results

In this work, generalized linear models on crash analysis system dataset are compared on various measures like plotting residuals, comparing the error metrics like root mean square error, mean square error and mean absolute error, comparing the goodness of fit test of models, etc. For comparing the goodness of fit of the models, 30 important features of are used which are selected using the recursive feature elimination method. Then the features are used in building the ordinary least squares(OLS) model. Then again using the same features poisson and negative binomial regression models are modelled(negative binomial model with default alpha value). Later the significance of the features is tested by comparing the p-values between the OLS model and the GLMs. Finally, the likelihood ratio test is performed for checking the best model among the generalized linear models.

Comparison of likelihood values of Poisson and Negative binomial Models: A regression model's log-likelihood value is a technique for determining the model's goodness of fit. The greater the log-likelihood number, the better the model matches the dataset. It can be seen from the figures 4.4.1 and 4.4.2, the log- likelihood value of negative binomial model is greater than the value of poisson model. For a particular model, the log-likelihood value might vary from $-\infty$ to $+\infty$

Generalized Linear Model Regression Results

Dep. Variable:	seriousInjuryCount	**No. Observations:**	617754
Model:	GLM	**Df Residuals:**	617724
Model Family:	Poisson	**Df Model:**	29
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1.5039e+05
Date:	Tue, 19 Apr 2022	**Deviance:**	2.2450e+05
Time:	19:37:59	**Pearson chi2:**	7.95e+05
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.03341
Covariance Type:	nonrobust		

Fig 4.4.1 Log-Likelihood of the Poisson model: -1.5039e+05

Generalized Linear Model Regression Results

Dep. Variable:	seriousInjuryCount	**No. Observations:**	617754
Model:	GLM	**Df Residuals:**	617724
Model Family:	NegativeBinomial	**Df Model:**	29
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1.4753e+05
Date:	Tue, 19 Apr 2022	**Deviance:**	1.8824e+05
Time:	19:39:09	**Pearson chi2:**	7.49e+05
No. Iterations:	10	**Pseudo R-squ. (CS):**	0.03162
Covariance Type:	nonrobust		

Fig 4.4.2 Log-Likelihood of Negative Binomial model: -1.4753e+05

V. NEURAL NETWORKS

5.1 Modelling using Neural Network Model:

Artificial neural networks (ANNs)[29,31,32] are remarkably similar to neurons in the human brain in terms of how they work. These ANNs undertake sophisticated mathematical calculations and produce a single output from several inputs. The activation function is a weighted input summation with bias as a determined function. Figure 5.1.1 is the illustration of the architecture of neural network for this work.

- Weights - These are the values that are multiplied to input and then added to form output. These also determine the effect of one neuron on another.
- Layer - As shown in the figure, the set of neurons sent as input to the network is called the input layer.
- Hidden layer - The layers where the network processes and learns from the data.
- Output layer - This layer produces the results.
- Activation function – This is the function applied to the output from the neurons.
- Bias - This is an additional term added to the linear equation. This parameter helps adjust the output corresponding to the weighted input sum.

Rectified Linear Unit Activation function:

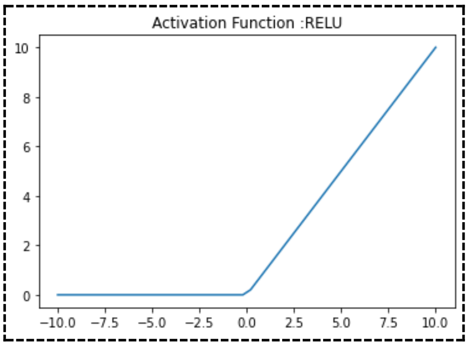

Fig 5.1.2 ReLu Activation

We use Relu activation as it is a regression problem and our scenario accepts only non-negative values. This activation function is mostly used in recent times. The formula for this function is f(x) = 0 for x < 0 , x for x>=0 and that is shown in the figure 5.1.2. The function's output varies from 0 to infinity. It is unbounded from above. This activation function is mostly applied in speech recognition and computer vision which uses deep neural networks.

Sequential Neural Network Architecture:

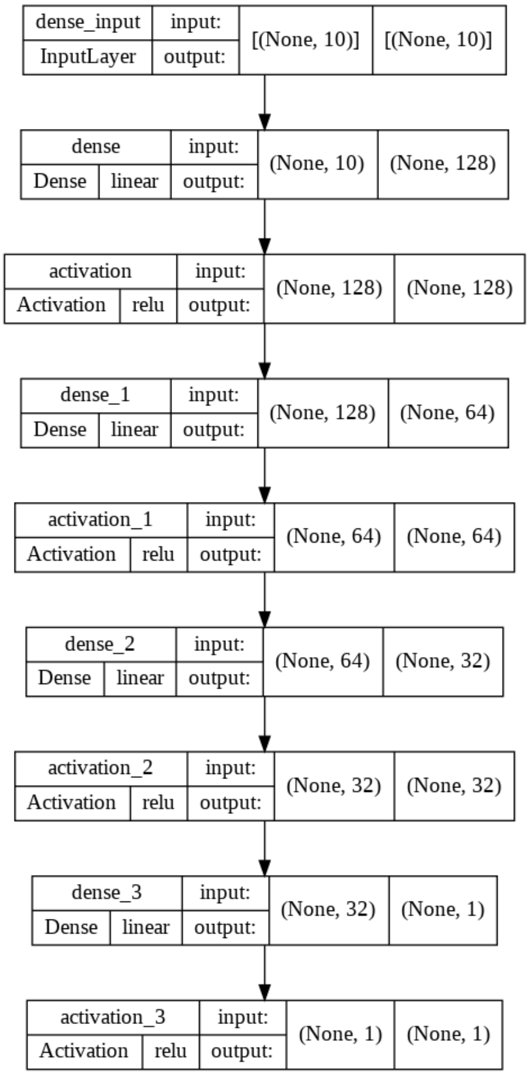

Fig 5.1.1: Architecture of Neural Network Model

The activation functions are responsible for introducing non-linearity. This is a mathematical function that transforms one input into another. The neural network is no better than a linear model if the activation function is not used. The output of neurons with a linear activation function is proportional to the input. Between input and output terms, there is a linear connection. The non-linear activation function is used to tackle various complicated issues. These activation functions give neural networks non-linear properties.

5.2 Neural Network model Results

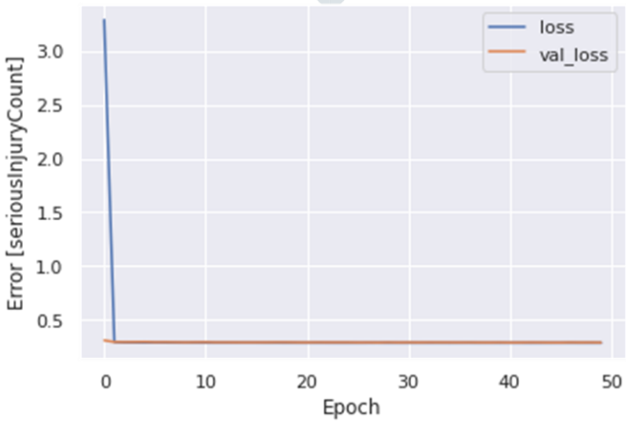

Fig 5.2.1 loss and Validation Loss

In this simple implementation of the sequential neural network model, we trained for 50 epochs with batch size 256, 0.01 being the initial learning parameter, and rmse being the loss. Table 5.2.1 gives us the parameters used in neural network model.

Activation Function	ReLu
Loss	RMSE
RMSE	0.290
MAE	0.120
Test loss and train loss difference	0.0030
Initial Learning Parameter	0.01
Epochs	50
Batch Size	256
Dense layers	3
No. of parameters in dense layer-1	1408
No. of parameters in layer-2	8256
No. of parameters in layer-3	2080
No. of parameters in layer-4	33
Total trainable parameters	11,777
Total number of parameters	11,777
Optimizer	Adam

Table 5.2.1 Parameters used in neural network model

Here from figure 5.2.1, it can be seen that there is no difference in loss and validation loss after the first few epochs. When it is compared with glm this neural network model took more time in training.

VI. CONCLUSION

For comparing the performance of various Regression models, the root mean square error and the mean absolute error metrics are considered as shown in table 6.1. Those two are the measures of differences between the actual and the predicted values of the model on test data. The absolute distance between the data and the regression predictions is calculated by **MAE** using the average of all observations. We use the absolute value of the distances to account for negative mistakes efficiently. You may also square the distance (**MSE)** to get a great result. Because of the nature of the power function, larger errors (or distances) are more essential in the measurement than smaller ones. The square root of the MSE error(**RMSE**) is then utilized to restore it to its original unit, but the property of punishing larger errors remains. We have the RMSE and MAE values of the Poisson, quasi-Poisson, Negative binomial, and Neural network models.

Model Used	**RMSE**	**MAE**
Poisson Regression	1.478	1.104
Quasi-Poisson Regression	1.121	1.003
Negative Binomial Regression	0.302	0.131
Neural Network Model	0.290	0.120

Table 6.1 Error Metrics values of various Models

Even though the neural network model is having slightly small error values, the Negative Binomial model is considered as it took very less time in building the model when compared to the neural network model.

ACKNOWLEDGMENTS

We dedicate this work to the Revered Founder Chancellor of Sri Sathya Sai Institute of Higher Learning, Bhagawan Sri Sathya Sai Baba. We acknowledge expert advice received from Mr. T.V.S Aditya, Actuarial Data Scientist.

REFERENCES

[1] et. al, R. K. S. "Study on Actuarial Science and Actuaries-The Pioneer of Insurance". International Journal of Advanced Science and Technology, Vol. 28, no. 20, Nov. 2019, pp. 1266 -81, http://sersc.org/journals/index.php/IJAST/article/view/4283.

[2]Zeileis, Achim, Christian Kleiber, and Simon Jackman. "Regression models for count data in R." Journal of statistical software 27.8 (2008): 1-25.

[3]Ismail, Noriszura, and Hossein Zamani. "Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models." Casualty Actuarial Society E-Forum. Vol. 41. No. 20. Spring, 2013.

[4]Byers, Amy L., et al. "Application of negative binomial modelling for discrete outcomes: a case study in aging research." Journal of clinical epidemiology 56.6 (2003): 559-564.

[5]Yirga, Ashenafi A., et al. "Negative binomial mixed models for analyzing longitudinal CD4 count data." Scientific reports 10.1 (2020): 1-15.

[6]Noll, Alexander and Salzmann, Robert and Wuthrich, Mario V., Case Study: French Motor Third-Party Liability Claims (March 4, 2020). Available at SSRN: https://ssrn.com/abstract=3164764 or http://dx.doi.org/10.2139/ssrn.3164764

[7]Huerta, E.B., Caporal, R.M., Arjona, M.A., Hernández, J.C.H. (2013). Recursive Feature Elimination Based on Linear Discriminant Analysis for Molecular Selection and Classification of Diseases. In: Huang, DS., Jo, KH., Zhou, YQ., Han, K. (eds) Intelligent Computing Theories and Technology. ICIC 2013. Lecture Notes in Computer Science(), vol 7996. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39482-9_28

[8]Shaaban, A.N., Peleteiro, B. & Martins, M.R.O. Statistical models for analyzing count data: predictors of length of stay among HIV patients in Portugal using a multilevel model. BMC Health Serv Res 21, 372 (2021). https://doi.org/10.1186/s12913-021-06389-1

[9]https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::crash-analysis-system-cas-data-1

[10]https://www.javatpoint.com/feature-selection-techniques-in-machine-learning

[11]Aha, D.W., Bankert, R.L. (1996). A Comparative Evaluation of Sequential Feature Selection Algorithms. In: Fisher, D., Lenz, HJ. (eds) Learning from Data. Lecture Notes in Statistics, vol 112. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-2404-4_19

[12]https://en.wikipedia.org/wiki/Poisson_distribution

[13]https://www.researchgate.net/publication/288022415_AN_APPLICATION_OF_CLAIM_FREQUENCY_DATA_USING_ZERO_INFLATED_AND_HURDLE_MODELS_IN_GENERAL_INSURANCE

[14]https://www.researchgate.net/publication/228856559_Spatial_modelling_of_claim_frequency_and_claim_size_in_non-life_insurance

[15]https://www.statology.org/poisson-distribution-assumptions

[16]Haghani S, Sedehi M, Kheiri S. Artificial Neural Network to Modelling Zero-inflated Count Data: Application to Predicting Number of Return to Blood Donation. J Res Health Sci. 2017;17(3):392. Published 2017 Sep 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7189957/

[17]Ridout, Martin, Clarice GB Demétrio, and John Hinde. "Models for count data with many zeros." Proceedings of the XIXth international biometric conference. Vol. 19. Cape Town, South Africa: International Biometric Society Invited Papers, 1998.

[18]Peiming Wang, A bivariate zero-inflated negative binomial regression model for count data with excess zeros,Economics Letters,Volume 78, Issue 3,2003,Pages 373-378,ISSN 0165-1765, https://doi.org/10.1016/S0165-1765(02)00262-8. (https://www.sciencedirect.com/science/article/pii/S0165176502002628)

[19]Chen, Xue-wen & Jeong, Jong Cheol. (2008). Enhanced recursive feature elimination. 429 - 435. 10.1109/ICMLA.2007.35.

[20]Schober, Patrick MD, PhD, MMedStat*; Vetter, Thomas R. MD, MPH† Count Data in Medical Research: Poisson Regression and Negative Binomial Regression, Anaesthesia & Analgesia: May 2021 - Volume 132 - Issue 5 - p 1378-1379 ,doi: 10.1213/ANE.0000000000005398

[21]Use and Extension of Count Data Models in the Determination of Relevant Factors for Claims in the Automobile Insurance Sector.Jose Antonio Ordaz, Maria del Carmen Melgar, M. Kazim Khan European Research Studies Journal, Volume XIII, Issue 4, 119-138, 2010 DOI: 10.35808/ersj/303

[22]Mackenzie J. Edmondson, Chongliang Luo, Md. Nazmul Islam, Natalie E. Sheils, John Buresh, Zhaoyi Chen, Jiang Bian, Yong Chen, Distributed Quasi-Poisson regression algorithm for modelling multi-site count outcomes in distributed data networks, Journal of Biomedical Informatics, Volume 131, 2022, 104097, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2022.104097. (https://www.sciencedirect.com/science/article/pii/S1532046422001137)

[23]Mackenzie J. Edmondson, Chongliang Luo, Md. Nazmul Islam, Natalie E. Sheils, John Buresh, Zhaoyi Chen, Jiang Bian, Yong Chen, Distributed Quasi-Poisson regression algorithm for modelling multi-site count outcomes in distributed data networks, Journal of Biomedical Informatics, Volume 131, 2022, 104097, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2022.104097.
(https://www.sciencedirect.com/science/article/pii/S1532046422001137)

[24]George Tzougas, Alice Pignatelli di Cerchiara, The multivariate mixed Negative Binomial regression model with an application to insurance a posteriori ratemaking, Insurance: Mathematics and Economics, Volume 101, Part B, 2021, Pages 602-625,

[25]Peng Shi, Emiliano A. Valdez, Multivariate negative binomial models for insurance claim counts, Insurance: Mathematics and Economics, Volume 55, 2014, Pages 18-29, ISSN 0167-6687,
https://doi.org/10.1016/j.insmatheco.2013.11.011.
(https://www.sciencedirect.com/science/article/pii/S0167668713001947)

[26]William L. Thompson, Gary C. White, Charles Gowan, Chapter 1 - Basic Concepts, Editor(s): William L. Thompson, Gary C. White, Charles Gowan, Monitoring Vertebrate Populations, Academic Press, 1998, Pages 1-42,
ISBN 9780126889604,
https://doi.org/10.1016/B978-012688960-4/50001-0.
(https://www.sciencedirect.com/science/article/pii/B9780126889604500010)

[27]Karen C.H. Yip, Kelvin K.W. Yau, On modelling claim frequency data in general insurance with extra zeros, Insurance: Mathematics and Economics, Volume 36, Issue 2, 2005, Pages 153-163, ISSN 0167-6687,
https://doi.org/10.1016/j.insmatheco.2004.11.002.
(https://www.sciencedirect.com/science/article/pii/S0167668704001301)

[28]Lluís Bermúdez, Dimitris Karlis, Bayesian multivariate Poisson models for insurance ratemaking, Insurance: Mathematics and Economics, Volume 48, Issue 2, 2011, Pages 226-236, ISSN 0167-6687,
https://doi.org/10.1016/j.insmatheco.2010.11.001.
(https://www.sciencedirect.com/science/article/pii/S0167668710001289)

[29]Gonca Buyrukoğlu, Selim Buyrukoğlu, Zeynal Topalcengiz, Comparing Regression Models with Count Data to Artificial Neural Network and Ensemble Models for Prediction of Generic Escherichia coli Population in Agricultural Ponds Based on Weather Station Measurements, Microbial Risk Analysis, Volume 19, 2021, 100171, ISSN 2352-3522,

https://doi.org/10.1016/j.mran.2021.100171.
(https://www.sciencedirect.com/science/article/pii/S235235222100013X)

[30]Hinde, John, and Clarice GB Demétrio. "Overdispersion: models and estimation." Computational statistics & data analysis 27.2 (1998): 151-170.

[31]Mitra, Sudeshna, and Simon Washington. "On the nature of over-dispersion in motor vehicle crash prediction models." Accident Analysis & Prevention 39.3 (2007): 459-468.

[32]Sharma, Sagar, Simone Sharma, and Anidhya Athaiya. "Activation functions in neural networks." towards data science 6.12 (2017): 310-316.