Universal Journal of Accounting and Finance 9(1): 1-14, 2021 http://www.hrpub.org
DOI: 10.13189/ujaf.2021.090101

Predictive Modeling of Insurance Claims Using Machine Learning Approach for Different Types of Motor Vehicles

V. Selvakumar[1,2,*], Dipak Kumar Satpathi[3], P. T. V. Praveen Kumar[3], V. V. Haragopal[3]

[1]Department of Mathematics, BITS - Pilani, Hyderabad, India
[2]Bhavan's Vivekananda College of Science, Humanities, and Commerce, Hyderabad, India
[3]Department of Mathematics, BITS -Pilani, Hyderabad, Telangana 500078, India

Received August 6, 2020; Revised December 16, 2020; Accepted January 20, 2021

Cite This Paper in the following Citation Styles

(a): *[1] V. Selvakumar, Dipak Kumar Satpathi, P. T. V. Praveen Kumar, V. V. Haragopal , "Predictive Modeling of Insurance Claims using Machine Learning Approach for Different Types of Motor Vehicles," Universal Journal of Accounting and Finance, Vol. 9, No. 1, pp. 1 - 14, 2021. DOI: 10.13189/ujaf.2021.090101.*

(b): *V. Selvakumar, Dipak Kumar Satpathi, P. T. V. Praveen Kumar, V. V. Haragopal (2021). Predictive Modeling of Insurance Claims using Machine Learning Approach for Different Types of Motor Vehicles. Universal Journal of Accounting and Finance, 9(1), 1 - 14. DOI: 10.13189/ujaf.2021.090101.*

Abstract The main objective of this research paper is to build an appropriate mathematical model that helps in forecasting third party claim amount for different categories of vehicles based on the chosen characteristics of the data. In actuarial research, predicting the insurance claim amount for different vehicle categories is a challenging task, and minimal empirical research studies were done to forecast the claims. In the present study, the annual time series historical data were collected for a period of 34 years. We had built the machine learning predictive models to modeling the claim amount with different categories of vehicles effectively. In this context, we exhibited the feasibility of using a statistical machine learning approach such as Linear regression Model, the Exponential Smoothing Model, autoregressive integrated moving average (ARIMA), artificial neural network (ANN), and hybrid ARIMA-ANN models to predict the various categories of vehicles claim amount. The data were analyzed, compared, and the empirical analysis showed that Artificial Neural Network is a better predictive model among the other time series models based on performance evaluation metrics RMSE and MAPE with lesser variance. Therefore, the machine learning approach for forecasting third party claim amounts will help the Insurance Companies in India to provide a better predictive model, which ensures better claims settlement and management for different categories of vehicles.

Keywords Linear Model, Box-Jenkins Model, Multilayer Perceptron, TRAINLM, Hybrid Model

1. Introduction

Motor Insurance is one of the most exciting branches of the insurance sector. In the year 1895, the third-party liability insurance policy was introduced in the Insurance field. The accidental damage for the four-wheelers was added to the policy in 1899. In India, the Motor Vehicles Act was passed in 1939. However, the provisions of compulsory third party (TP) insurance were introduced in the act only on 1st July 1946. For Insurance purposes, the Motor business in India, the motor vehicles are broadly categorized into Motorized two-wheelers, Private cars, Commercial four-wheelers, Commercial three-wheelers, Special types of vehicles. The motor insurance is also classified into third-party liability insurance claims and comprehensive package policy claims. In current years, the insurance sector of a commercial motor vehicle has heavily suffered due to third party damage losses based on some of the issues faced by the motor insurance industry of India such as there is no cooperation of insured for declaring claim settlement, the viability of motor insurance companies and there is no flexibility of law for

dealing of the claim.To effectively build a suitable mathematical model for the third-party claim amount to help insurance companies with a proper claim settlement with accuracy

In this context of TP claim data modeling, most researchers have neither evaluated the data nor modeled the existing data. In the literature, very few researchers studied information related to marine insurance, fire insurance, health insurance, etc., by predicting the number of claims and not used any of the other methodologies to forecast the claim amount, but there was a broad application of ANN model in different fields. A study [1] reveals that the ANN model is more potent in forecasting power distribution data than the time series models. Based on the comparative review of different models [2], the ANN models outperform the other traditional models in predicting the demand with greater accuracy. The ANN predictive efficiency is better than GARCH models in forecasting the stock exchange rate [3]. Another stock market exchange study concluded that ANN is an appropriate model for forecasting capital markets such as stock and currency [4]. Based on forecasting, electric power consumption in educational institutions suggested that the ANN model's developed structure performed good prediction [5]. Comparing different models [6] showed the hybrid model backpropagation is better than both the BPNN and ARIMA in all criteria for forecasting the sales. In the application of weather forecasting [7], the ANN approach with increased hidden layers predicts the maximum temperature with higher accuracy in a year. Another study revealed that the hybrid model with a combination of ARIMAX and NN showed better forecasts than individual forecasting models [8]. A survey of predicting stock market returns showed that non-linear models are better for forecasting the returns in emerging and frontier markets [9]. For predicting the market demand [10] and also improving the academic performance of the institution [11], the ANN model gives better accuracy as compared with other advanced models [12]. In a study on predicting the electricity demand in Thailand, the ANN model showed a more significant prediction [13]. By using the Multi-layer perceptron architecture ANN model machine learning approach, the traffic flow in Morocco country was predicted [14] with accuracy. Another study [15] predicted river runoff using ANN with accuracy. Apichottanakul et al. [16] have forecasted the market share of Thai rice. Other studies [17] related to the feed mix industry also showed that the production rate and dust level enhance the mill's capability by using ANN predicted well. Some research showed that the hybrid ARIMA-ANN model improved the accuracy of forecasting the resource usage in server virtualization as compared to ARIMA and ANN separately [18]. Another study also suggested that the hybrid ARIMA-ANN model has the best model for forecasting Indian Robusta coffee projection [19]. In recent years, machine learning techniques are developed to predict the financial time series data with greater accuracy [20]. A study on spine surgery, a machine learning predictive model, is designed to improve risk adjustment with greater efficiency [21]. In health care insurance claims, the Recurrent Neural Network (RNN) model shows better performance than other regression models [22]. In another study related to health insurance claim data, a machine learning predictive regression model LASSO was developed to formulate a population health management in Japan [23]. A study to forecast the stock exchange by applying machine learning techniques such as stacking and blending gives better prediction than bagging and boosting [24]. In another study, a machine learning approach was applied for modeling volatility in the pricing of deposit insurance [25]. For predicting accidental claims using telematics data, the logistic regression showed better prediction than the XGBoost machine learning algorithm [26]. However, the previous research on insurance claim modeling is very little, and few authors have considered the ARIMA model for prediction fire insurance, property insurance, and health care insurance. In India's motor insurance sector, there is no substantial amount of research study has been done, which remains a motivating factor for us to build appropriate forecasting models.

2. Research Methodology

2.1. Data Used for Research

From different Public Insurance companies of India, the secondary data are collected for distinctive 34 years from 1985 to 2018. The secondary data consists of 108 column variables and 9,62,689 row values (i.e., approximately One hundred three million nine hundred seventy thousand data points). We have studied the third-party claim amount variable for different categories of motor vehicles from these data points. Out of this 9,62,689 third party claim data set, 1,32,685 consists of Two-wheeler claims, 2,05,420 consists of Private car claims, 4,51,978 consists of Commercial four-wheeler claims, 1,25,487 consists of Commercial three-wheeler vehicle claims, and 47,119 belongs to Special type of vehicle claims. Further, we have divided each category of vehicle claim data randomly into 70% for training and 30% for the testing for fitting the best models using the machine learning approach. We now discuss the techniques of model building, as explained below.

2.2. Time Series Models

For the historical TP claim data set for various vehicles, we have applied a simple linear model, ARIMA, exponential Smoothing, ANN, and the combination of

linear & non-linear domain, a hybrid model. We then predicted the claim data for all the models and compared them with the statistical metrics such as Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The five forecasting modeling techniques used are as follows:

2.2.1. Linear Model

A simple linear regression model is a forecasting technique for predicting the TP claim amount by a given predictor variable time. The fitting of a simple linear model with an error term for predicting the claim amount based on historic data is provided by

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1}$$

Where X_i is a time variable in a yearly unit; Y_i is a TP claims amount; e_i is a residual error term

2.2.2. Exponential Smoothing

The exponential smoothing model uses the weighted average of past data to forecast the future claim amount when time-series data don't have any seasonality or trend and only have a level. This model will be represented as an equation to forecast the future claim amount is given as follow

$$\hat{y}_{(i+1)} = \alpha\{ y_i + (1-\alpha) y_{i-1} + (1-\alpha)^2 y_{i-2} + \cdots \} \tag{2}$$

$\hat{y}_{(i+1)} = \alpha\{y_i + (1-\alpha)y_{i-1} + (1-\alpha)^2 y_{i-2} + \cdots\}$ Where α is the smoothing parameter, $0 \le \alpha \le 1$; y_i is the third-party claim amount $\hat{y}_{i+1}$ is the forecast value of the claim amount.

2.2.3. ARIMA (p, d, q)

For third-party claim data for different vehicles, we have fitted ARIMA *(p, d, q)* model. In order to provide an ARIMA model efficiently, the data needs to be stationary. If the original data is non-stationary, then reconstruct it to stationary by differentiating the original series, then it will be modeled by ARIMA (*p, d, q)*. The autocorrelation function (ACF) and partial autocorrelation function (PACF) are evaluated, and it suggests the most appropriate ARIMA model for the claim data. The Box-Jenkins modeling approach illustrates four iterative steps: Identification of Model, Estimation of Parameters, Diagnostic Checks & Validation, and Model Forecasting. The mathematical equation of Box and Jenkins [27], ARIMA *(p, d, q)* process can be expressed as AR (p) I (d) MA (q) in terms of the backward shift operator:

$$\begin{aligned}(1 - \alpha_1 B_1 - \alpha_2 B_2 - \ldots\ldots\ldots - \alpha_p B_p)(X_t - \mu) = \\ (1 + \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_q B_q) e_t\end{aligned}$$

$$\begin{aligned}(1 - \alpha_1 B_1 - \alpha_2 B_2 - \ldots\ldots\ldots - \alpha_p B_p)(X_t - \mu) = (1 + \beta_1 B_1 + \\ \beta_2 B_2 + \cdots + \beta_q B_q) e_t\end{aligned} \tag{3}$$

Where B is describing the process of differencing, which is given by

$$y_t' = y_t - y_{t-1}$$

$$y_t' = y_t - B y_t$$

$$\Rightarrow \quad y_t' = (1 - B) y_t$$

In this context, we have constructed different ARIMA models to know which model fits the data well. For this, various statistical characteristics such as the RMSE, MAPE, Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC) are computed and compared for all vehicles to identify the best model.

2.2.4. Artificial Neural Network (ANN)

ANN is a mathematical or computational model of the machine learning technique interconnected with many neurons functioning together to resolve many complicated problems. The interconnection of artificial neurons, which emulates the function of human brains to solve scientific problems. Based on their interconnections, neural network models are developed. These networks are generally classified as a single layer and multi-layer perceptron model with feed-forward or feedback propagation. In the feed-forward network propagation model, the signals are moved from one neuron to another in a forward direction.

In many applications, the multi-layer feed-forward neural network (FFNN) model can predict the time series data. The multi-layer perceptron network consists of a multi-layer, an input layer, hidden layers, and an output layer. The model's validity is determined by neural network structure, methods of training or algorithms, and activation functions. Figure 1 shows the FFNN model structure with the input layer and the hidden layer, consisting of neurons.

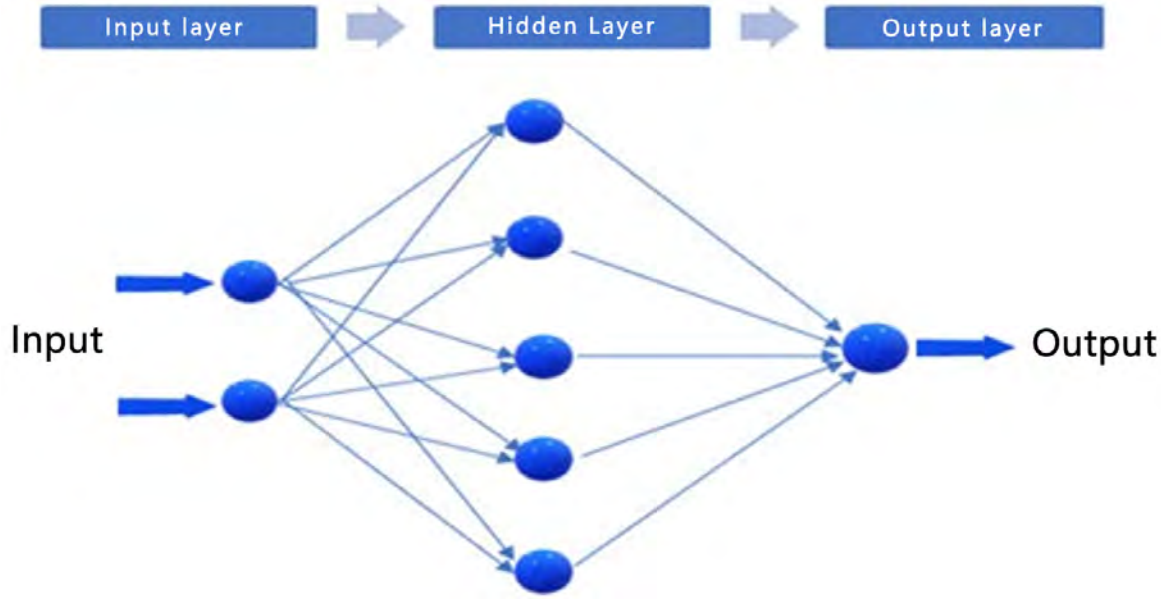

Figure 1. Neural Network Architecture

The mathematical relationship between the input $(y_{t-1}, y_{t-2}, \dots\dots, y_{t-p})$ and the output (y_t) for MLFFN can be written as

$$y_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i f(\sum_{i=1}^{n} \beta_{ij} y_{t-1} + \beta_{0j}) + e_i y_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i f(\sum_{i=1}^{n} \beta_{ij} y_{t-1} + \beta_{0j}) + e_i \quad (4)$$

Here 'α_i' is a weight from the hidden to output nodes, and '$\beta_{.j}$' weight from the input to hidden nodes; n and m refers to the number of Input nodes and number of hidden nodes; 'f' is the sigmoidal activation functions;

The sigmoidal function can be mathematically expressed as follows:

$$f(x) = \frac{1}{1+e^{-ax}} \quad (5)$$

2.2.5. Hybrid Model

In much real-world time series forecast modeling and as per the literature, neither ANN nor ARIMA is suitable for all situations because the time series contains both linear domain and non-linear domain problem structures. For a linear domain, the ARIMA model is ideal, whereas ANN is suitable for a non-linear domain. Both models have achieved success in their domains. Therefore, Zhang [28] proposed the mixed-use of linear and non-linear domains in a suitable methodology, a hybrid ARIMA-ANN model. In this approach, Zhang combined both ANN and ARIMA separately for modeling linear and non-linear components to evaluate accurate forecasting.

According to the Model, we have

$$y_i = L_i + N_i \quad (6)$$

Where y_i Stand for the third-party damage claim amount. L_i and N_i denote the linear and non-linear components of time series data.

According to Zhang, the ARIMA is fitted to the original series to model the linear part, and the corresponding residuals from the linear part contain only non-linear components that will be fitted by the ANN model. The residuals obtained from the ARIMA model are given by $e_i = y_i - \hat{L}_i$, where $\hat{L}_i$ stands for forecast value from ARIMA. Then, model the residuals using ANN that will capture the non-linear pattern of the TP claim data. Using n inputs, the ANN model for residuals will be of the form

$$e_i = f\{e_{i-1}, e_{i-2}, e_{i-3}, \dots\dots\dots e_{i-n}\} + \varepsilon_i \quad (7)$$

Where f represents the non-linear function obtained by the ANN and ε_i represents the random error. The forecasted value obtained from ANN equation (2) denoted as $\hat{N}_i$ then the hybrid ANN-ARIMA forecast to the time series data is obtained as

$$\hat{y}_i = \hat{L}_i + \hat{N}_i \quad (8)$$

The hybrid model is now applied to third-party claim data to check for better forecasting accuracy than ARIMA and ANN by comparing RMSE and MAPE.

2.3. Performance Criteria

The performance criteria such as Mean Absolute Percentage (MAPE) and Root Mean Square Error (RMSE) are shown in equations (9) and (10).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_t)^2} \quad (9)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| 100 \frac{y_i - \hat{y}_t}{y_t} \right| \quad (10)$$

3. Experimental Evaluation

The secondary dataset consists of 962119 third-party overall claim samples (all categories) of vehicles from a public insurance company. Further, these are categorized into 132685 two-wheeler claims, 205420 private cars, 451978 claims are Four wheelers, 124917 claims are three-wheelers, and 47119 are claims of a special type of vehicles. For this categorized data, we modeled the claim amount for the data. We evaluated the five different time series techniques empirically: Linear Model, Exponential Smoothing model, ARIMA model, ANN model, and hybrid ARIMA-ANN model to find which of these models fits well. The data analytics are analyzed by using STATGRAPHICS Version 18.1.12 and MATLAB 2019b Version.

4. Results and Discussion

4.1. Results of Linear Models

This article generates third-party claim forecasting models of the various types of motor vehicles by using three traditional time series models, ANN, and the hybrid ARIMA-ANN model based on secondary data. This study modeled exponential smoothing using STATGRAPHICS, while all other models are built using MATLAB. Then, all three evaluations are compared by the model's predicted values, and with actual values, the characteristics of RMSE and MAPE are estimated. From equation (1), we have fitted the generalized linear model, and it is observed that the coefficients $\beta_{0 \text{ and }} \beta_1$ are significant. Thus, by applying the coefficients β_0 and β_1 in equation (1), the linear model is fitted for different vehicle categories to predict the values. From Fig.2, it is observed that the claim data do not display any clear trends.

4.2. Results of Exponential Smoothing Models

From equation (2), we have identified the appropriate optimal smoothing constant α by using the trial and error method for the exponential smoothing method. For that, twelve experimental trials are performed with various smoothing parameters from 0.1 to 0.95 for each category of vehicles. We have generated the exponential smoothing model based on equation (2) with parameter α ($0 \leq \alpha \leq 1$). From Fig.3, it is observed that the minimum performance accuracy measures such as MAPE, MSE, RMSE are obtained at a larger value of the α (> 0.93) because the actual and forecasting TP claim amounts are fluctuating rapidly for all categories of vehicles. Also, it is clear that the smoothing constant α increases the performance measures of accuracy decrease.

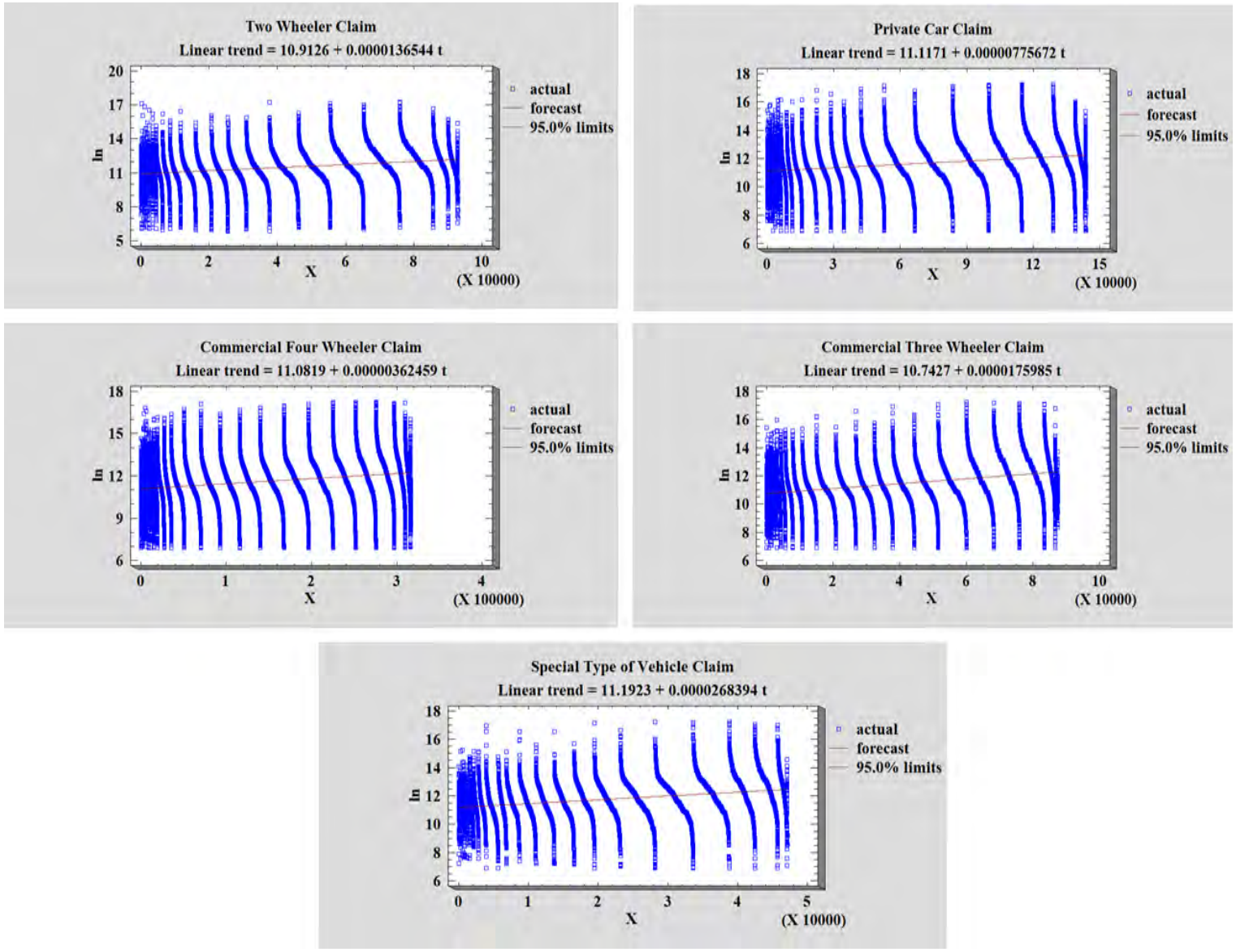

Figure 2. Linear Model plot for different types of Vehicle Claims

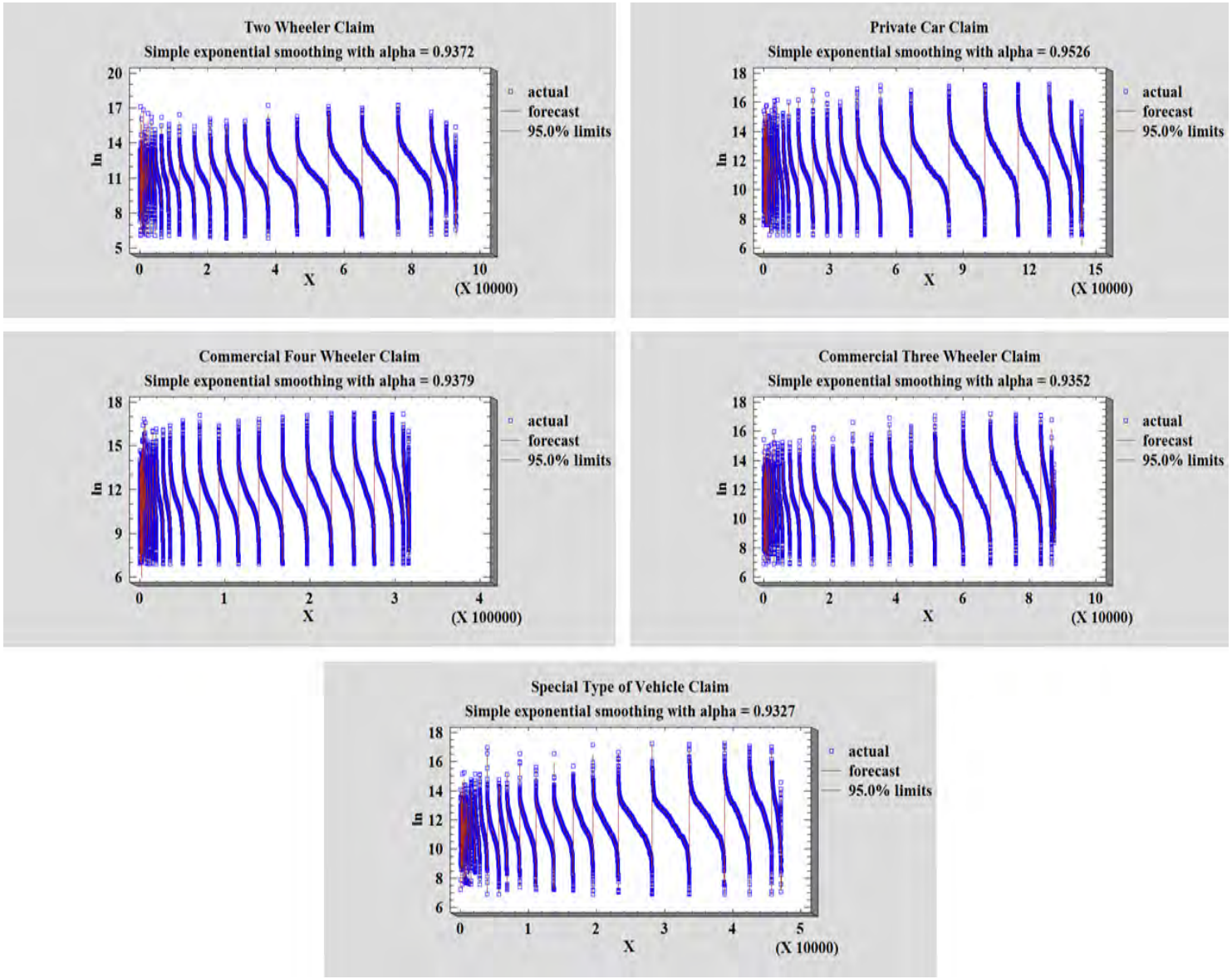

Figure 3. Exponential Smoothing M0odel plot for different types of Vehicle Claims

4.3. Results of ARIMA (p, d, q) Models

Based on equation (3) Box Jenkins methodology, the ARIMA models were built through the iterative building process, such as identifying the model, estimation, validation, and diagnostic checking. We first checked whether the time series data set for different categories of vehicles are stationary or not. We applied the Augmented Dickey-Fuller test for the original data set; the results concluded that the TP claim data of all categories of vehicles were stationary (i.e., $p < \alpha$). Now, the TP claim data for all types of vehicles will be modeled by ARIMA (p, d, q).

From TABLE 1, we have computed the best possible feasible ARIMA models and their performance criteria AIC, RMSE, MAPE, and BIC values. By comparing all the fitted ARIMA models based on their performance criteria, ARIMA (1, 0, 2) model fits well with relatively smaller performance criteria values mentioned above to forecast the TP Claim Amount for Two-wheelers, Private Cars, Commercial four-wheelers, Commercial Three Wheelers and Special type of Vehicles.

4.4. Results of Artificial Neural Network Models

By using Machine learning methodology, the analytical process was performed by considering the ANN procedures using MATLAB 2019b version workspace and computed the neural network technique evaluated. Out of the total samples, 70% of samples are selected randomly for training, 15% of them are chosen randomly for validating the network model, and the remaining 15% of them are the utterly independent test of network generalization.

We have selected the number of neurons in the hidden layer based on the desired performance using the TRAINLM method. By using the trial and error method, the performance of the network can be checked, and it has to be retrained if the performance is not satisfied. Finally, the TRAINLM method predicts the values based on Mean Square Error (MSE) for different vehicle categories.

Table 1. Performance Criteria values for the ARIMA *(1, 0, 2)* Models for different Categories of Vehicles

Statistics	Two Wheeler	Private Car	Commercial Four Wheeler	Three Wheeler	Special Type of Vehicle
RMSE	0.16705	0.12432	0.09060	0.15552	0.32744
MAE	0.00625	0.00381	0.00187	0.00598	0.02909
MAPE	0.05103	0.02993	0.01476	0.04778	0.23585
AIC	-3.5788	-4.1697	-4.80254	-3.7218	-2.2325
BIC	-3.57850	-4.1695	-4.80244	-3.7215	-2.23089

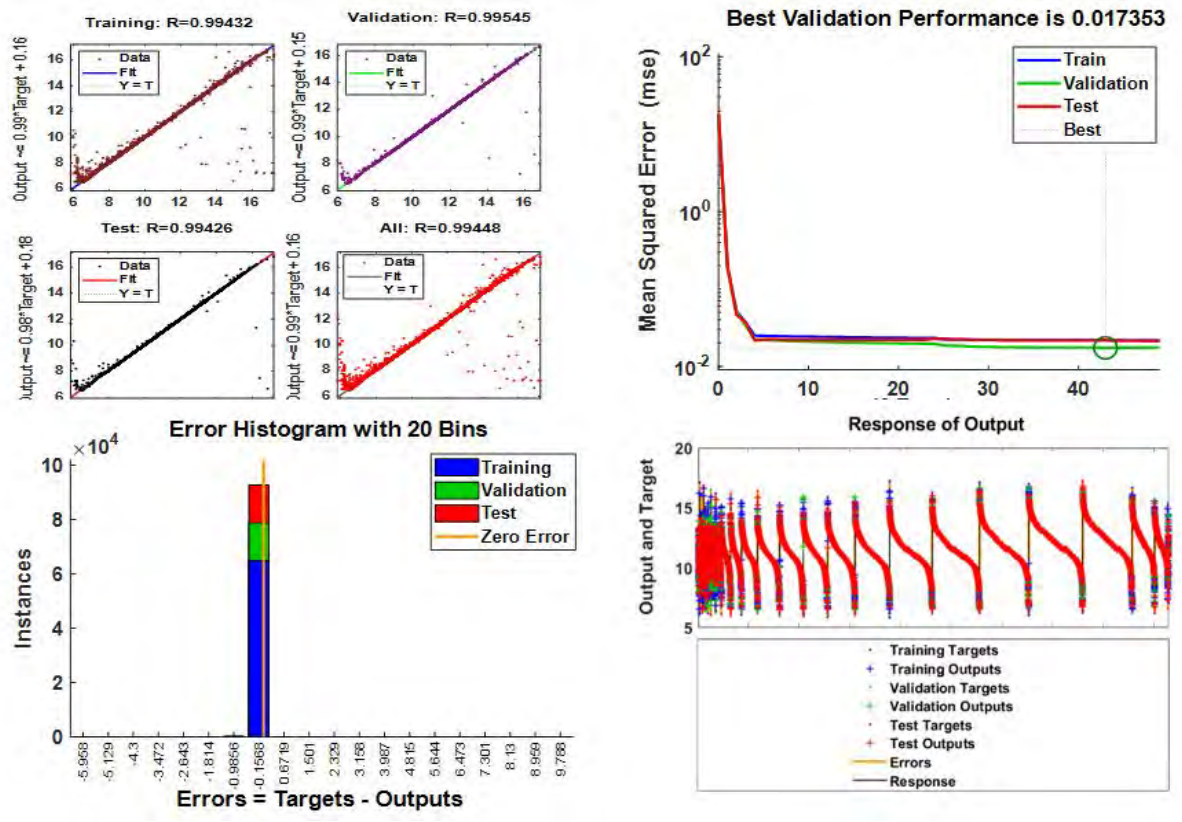

Figure 4. Regression plot, performance plot, Error histogram and Response plot for Two-wheeler Claims Using the ANN model

From Fig.4, the Regression plot for training, validation, and test claims data sets with respect to two-wheeler claims fall along a 45-degree line, and also R > 0.9941 for each case indicates the predicted regression plot showed a good fit. It also suggests that the network outputs are equal to the targets. The performance plot predicts the best performance validation RMSE as 0.017353. Also, the error histogram for training, validation, and testing data are represented by blue, green, and red bars gives an idea about the indication of outliers. It is observed that the error is zero; the prediction was successful.

From Fig. 5, the regression plots for private cars show a perfect fit with R > 0.997 for training, validation, and testing data set. It is also showing the best validation performance RMSE as 0.011243 with error histograms; the error is zero, which indicates that the predicted model is perfect.

From Fig. 6, the regression plots for the Commercial four-wheeler show a perfect fit in each case R values are above 0.998, with the resulting best validation performance RMSE as 0.005305, and the error is zero, which indicates that the predicted model is perfect.

From Fig. 7, the regression plots of ANN show a perfect fit with R-value is higher than 0.995 in each case, with the resulting best performance RMSE as 0.021751 for Commercial Three-wheeler, and the error is zero. It indicates that the predicted model is perfect.

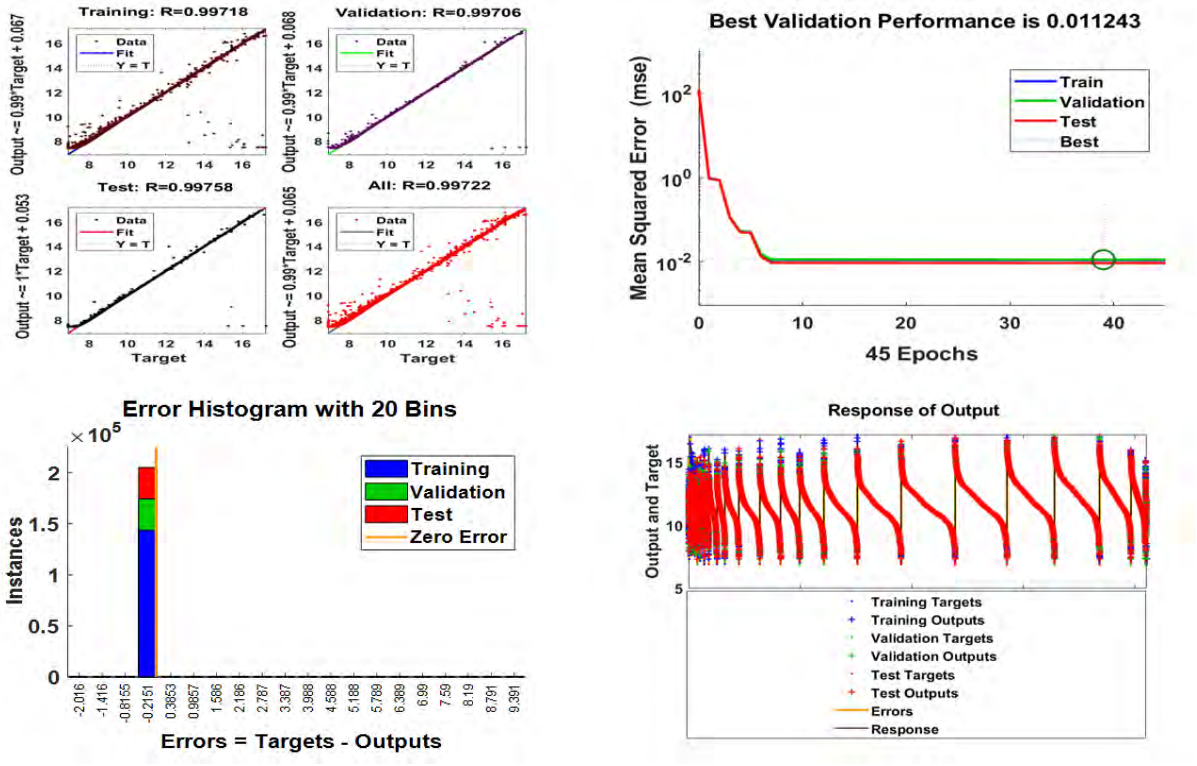

Figure 5. ANN model - Regression plot, performance plot, Error histogram and Response plot for Private Car Claims

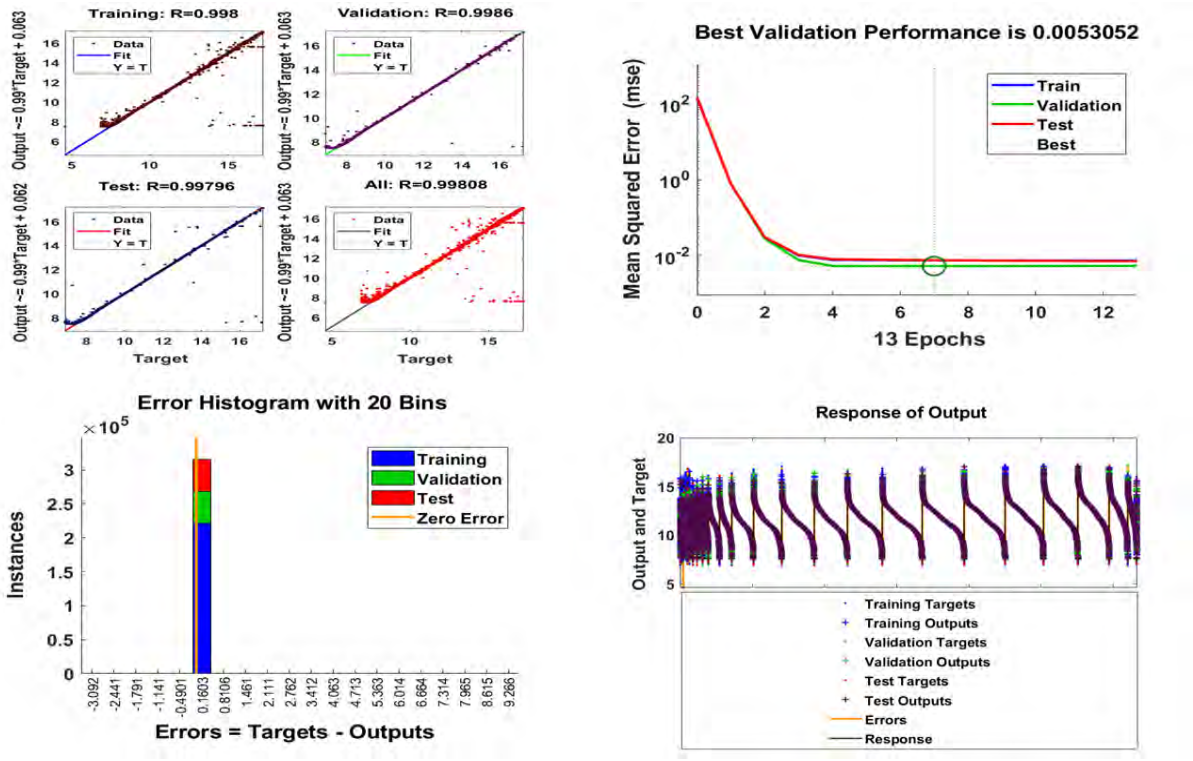

Figure 6. ANN model - Regression plot, performance plot, Error histogram, and Response plot for Commercial Four-wheeler Claims

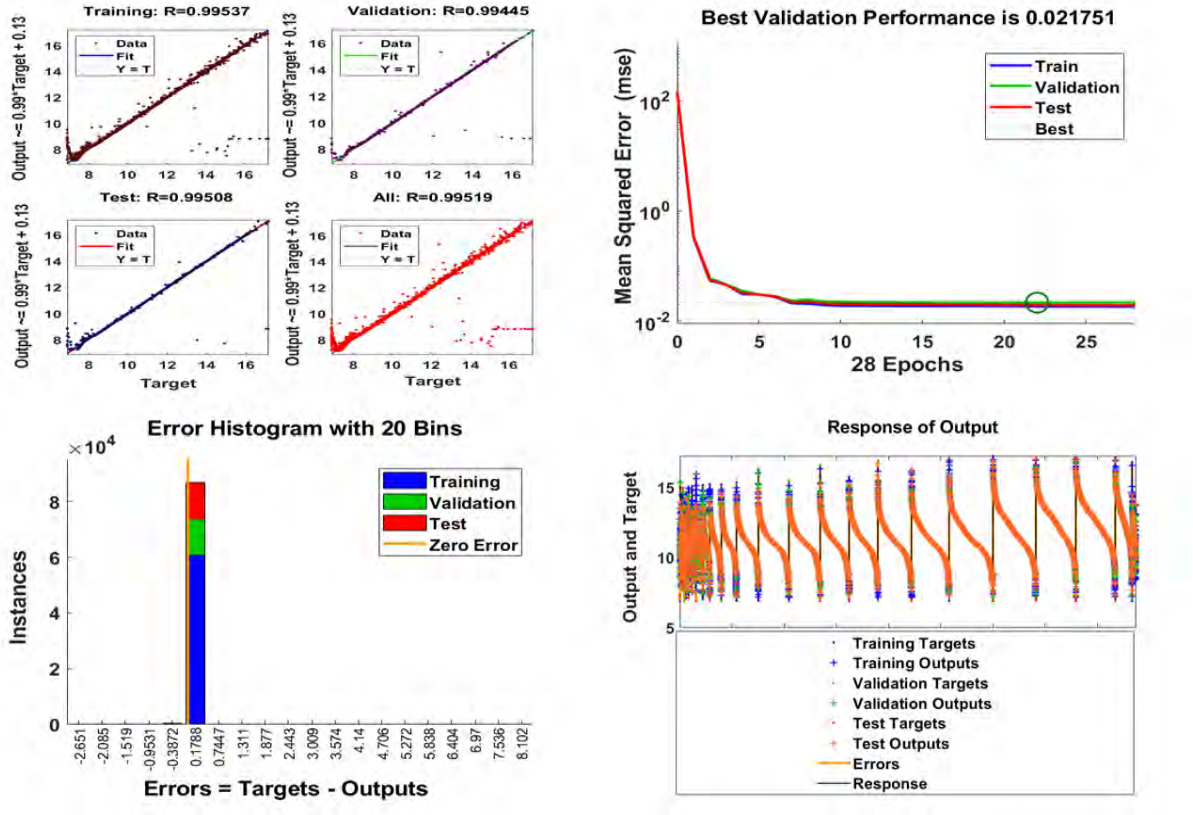

Figure 7. ANN model - Regression plot, performance plot, Error histogram and Response plot for Commercial Three-wheeler Claims

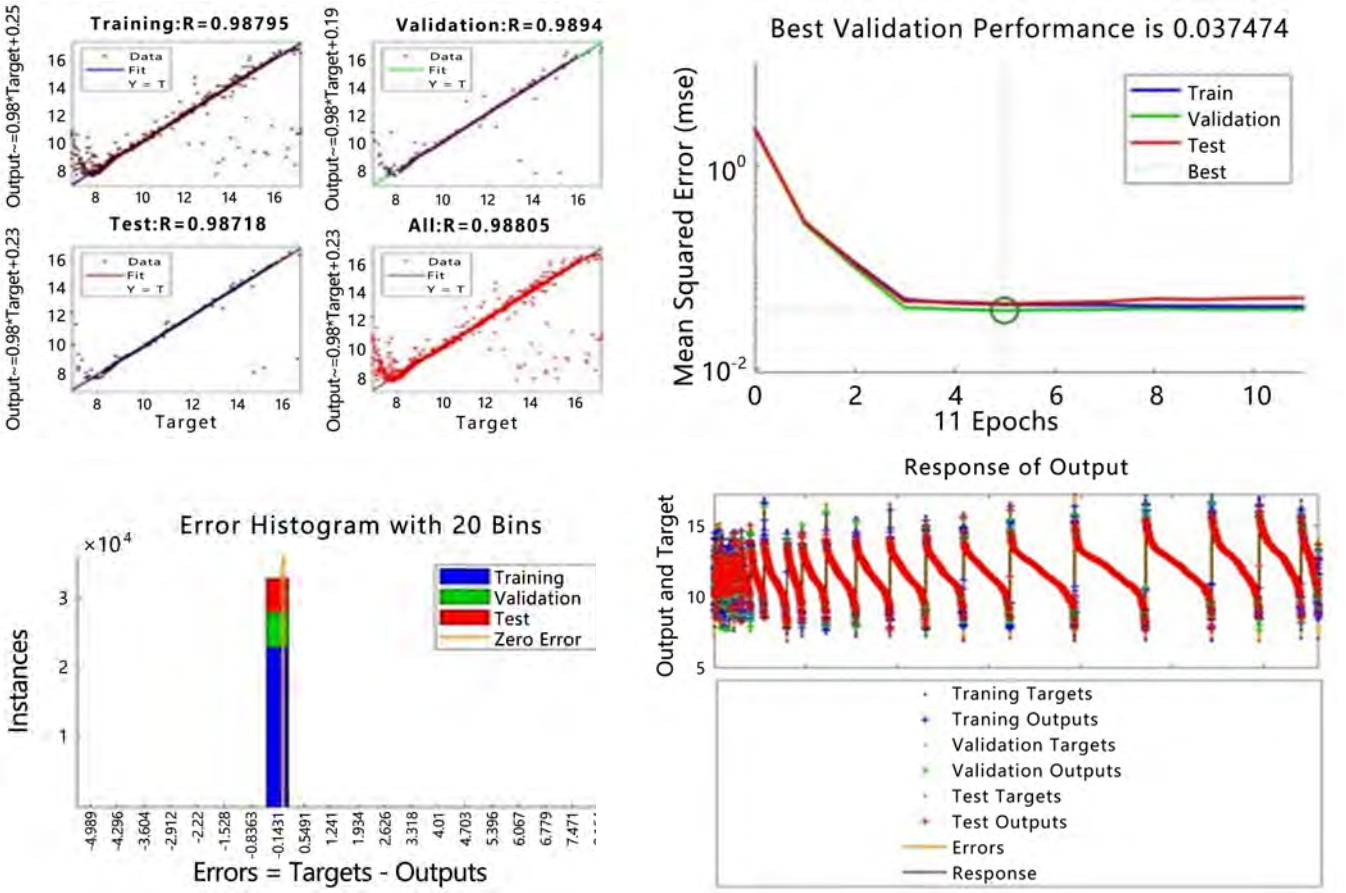

Figure 8. ANN model - Regression plot, performance plot, Error histogram and Response plot for Special type of vehicle Claims

From Fig. 8, it is observed for a special type of vehicle, the regression plots of ANN show a perfect fit with R values are above 0.998 for training, testing, and validation, with the resulting best performance RMSE as 0.037474, and the error histogram showed the error is zero.

From the above results based on all vehicle categories, the MLFF neural network with a sigmoid transfer function in the hidden layer and a linear function in the output layer prediction was accurate with less RMSE.

4.5. Results of Hybrid ARIMA-ANN Model

In this model, first, we have fitted the ARIMA Model for original data for all categories of vehicles, which is obtained from equation (3). The residuals obtained from the ARIMA model shows non-linearity (equation 7). Then, we have applied the ANN technique, MLFF neural network, to the ARIMA residuals by using the Levenberg-Marquardt algorithm.

Finally, according to G. Peter Zhang [28] equation (8), the hybrid model's predicted values are estimated by adding the predicted values of ARIMA and ANN. Also, we have evaluated the various performance measures of the hybrid model for different categories of vehicles.

4.6. Evaluation of Forecast Accuracy

From the literature review of a few authors suggested that ARIMA is an accurate model to predict the data with trend and seasonality [29,30]. Many of them suggested mixed use of the linear and non-linear hybrid model is a reliable model for forecasting the financial data.

In this study, we compared all five forecasting models, such as the Linear Model, Exponential Smoothing, ARIMA, ANN, and hybrid models for various categories of motor vehicles for TP claim accuracy.

Table 2. Comparison of forecast accuracy between Models for different Vehicle Categories

Vehicle Category	Model	Training Data Set		Testing Data Set		Training Data	Testing Data
		RMSE	MAPE	RMSE	MAPE		
Two-Wheeler Claim	Linear Model	1.33406	9.22843	1.32864	9.18160	92880	39805
	Exponential Smoothing	0.16722	0.04949	0.23507	0.10833		
	ARIMA	0.16705	0.05103	0.23469	0.10994		
	ANN	0.14527	0.04204	0.18714	0.10547		
	Hybrid	0.16660	0.12312	0.23040	0.28109		
Private Car Claim	Linear Model	1.352	9.3685	1.36274	9.4498	143794	61626
	Exponential Smoothing	0.1244	0.02897	0.18706	0.0672		
	ARIMA	0.1243	0.02993	0.18692	0.0675		
	ANN	0.11755	0.02697	0.16268	0.05029		
	Hybrid	0.12391	0.02975	0.18476	0.21816		
Commercial Four-Wheeler	Linear Model	1.32918	9.37124	1.32867	9.37040	316385	135593
	Exponential Smoothing	0.09062	0.01469	0.13145	0.03129		
	ARIMA	0.09060	0.01476	0.13132	0.03146		
	ANN	0.08489	0.01202	0.11985	0.06080		
	Hybrid	0.08793	0.06663	0.13048	0.09672		
Commercial Three-Wheeler	Linear Model	1.34259	9.33241	1.35022	9.36136	87742	37745
	Exponential Smoothing	0.15565	0.04702	0.22814	0.10470		
	ARIMA	0.15552	0.04778	0.22791	0.10457		
	ANN	0.13832	0.08848	0.17348	0.17451		
	Hybrid	0.15212	0.11847	0.21883	0.20914		
Special type of vehicle	Linear Model	1.28410	8.90154	1.29360	9.01618	32984	14135
	Exponential Smoothing	0.23366	0.21165	0.32799	0.24036		
	ARIMA	0.32744	0.23585	0.32744	0.23585		
	ANN	0.20455	0.17511	0.26450	0.22598		
	Hybrid	0.23058	0.28233	0.30908	0.38381		

From Table 2, the exploratory results showed the performance criteria index of RMSE and MAPE for different categories of vehicles with other models such as Linear Model, Exponential smoothing, ARIMA, and hybrid model for both trained and tested data set. Also, both trained and tested data indicated that the ANN model was better compared to other models. Out of five vehicle categories of insurance claims, the ANN model produced an optimal forecast for all vehicle categories. Another exciting conclusion is that the hybrid model and traditional time series models ARIMA, exponential smoothing has not become abortive; they are still handy for forecasting because the performance criteria values are less, as shown in Table 1. These comparative results showed that the ANN model yields a more accurate forecast than any other model for different categories of vehicles with lesser RMSE and MAPE.

5. Conclusions

In recent years, the growing demand for a motor insurance segment, due to the inherent risk factor, the unpredictable occurrence of the motor insurance claim, and the complex nature of claim data, accurately predicting third party claim amount for various categories of vehicles can help the Motor Insurance companies of India to provide a better customer-centric forecasting model which ensures better claims settlement and management. In this context, we have applied a time series modeling technique such as Linear model, Exponential Smoothing, ARIMA, ANN, and hybrid ARIMA-ANN model to see which of these models are better for forecasting the claim amount and calculated the performance measure: RMSE and MAPE for all models.

From the data analytics performed by time series analysis, ARIMA (*1, 0, 2)* model fits well for all categories of vehicles by using the Box-Jenkins approach based on the respective performance criteria that are AIC, BIC, MAPE, and RMSE. Comparing the performance criteria of different models based on trained and tested datasets for each category of vehicles, the ANN model yields a more accurate forecast compared to all other models because of its non-linear claim data set.

Finally, this exploratory analysis concluded that the ANN models outperform the other predicting models in forecasting the insurance claim amount with greater accuracy. As per the literature studies, the predictions are performed by a hybrid model, which fits well with the data. This study reveals an interesting fact that the hybrid model was not found to be appropriate for any category of vehicle type because the time series contains non-linear data set. Thus, the ANN model is suggested as the best model for forecasts of any vehicle category compared to the linear model, exponential smoothing, ARIMA, and hybrid modeling. This exploratory machine learning analytics approach would help motor insurance companies deal with the uncertain occurrence of claims and ensure the Claims' easy disbursement. Also, this Artificial Neural Network customer-centric forecasting modeling approach helps the insurance companies to provide better claim settlement and management.

REFERENCES

[1] Ronna chai Ch, Supaporn B, Chanintorn T, Nittaya K, and K Kerdprasop. Artificial Neural Networks and Time Series Models for Electrical Load Analysis, Proceedings of the International Multiconference of Engineers and Computer Scientists, Volume 1, 271-279, 2016.

[2] Shubham B and Arvind J. Development of ANN Models for Demand Forecasting, American Journal of Engineering Research, Vol. 6, No.12, 142-147, 2017.

[3] Tamal Datta Chaudhuri and Indranil Ghosh. Artificial Neural Network and Time Series based Approach to Forecasting the Exchange Rate in a Multivariate Framework, Journal of Insurance and Financial Management, Vol.1, No.5, 92-123 2016.

[4] Parsinejad Shahbaz, Bagheri Ahmad, Ebrahimi Atani Reza and Javadi Moghaddam Jalal. Stock Market Forecasting Using Artificial Neural Networks, European Online Journal and Natural and social Sciences, Vol.2, 2013.

[5] Samsher Kadir Sheikh and M G Unde. Short Term Load Forecasting using ANN Technique, International Journal of Engineering Sciences & Emerging Technologies, Vol. 1, No.2, 97-107, 2012.

[6] Hani Omar, Van Hai Hoang and Duen-Ren Liu. A Hybrid Neural Network Model for Sales Forecasting based on ARIMA and Search Popularity of Article Titles, Computational Intelligence and Neuroscience, Vol.1, 2016.

[7] Kumar Abhishek, M P Singh, Saswata G, Abhishek A. Weather forecasting model using Artificial work, Procedia Technology, Elsevier, Vol.4, 311-318, 2012.

[8] Suhartono, Rahaya S P, Prastyo D D, Wijayanti D G P, Juliyanto. Hybrid model for forecasting time series with trend, seasonal and salendar variation patterns, IOP Conference Series: Journal of Physics, Vol. 890, No.012160, 2017.

[9] Mallikarjuna, M., Rao, R.P. Evaluation of forecasting methods from selected stock market returns, Springer Open, Financial Innovation, vol. 5, No.40, 2019.

[10] E.T. Lau, L. Sun and Q Yang, Modelling, prediction and classification of student academic performance using artificial neural networks, Springer Nature, SN Applied Sciences, Vol.1, No.982, 2019.

[11] Z. Ying and X. Hanbin. Study on the Model of Demand Forecasting Based on Artificial Neural Network, Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, IEEE, Hong Kong, 382-386, 2010.

[12] Karin Kandananond. A comparison of various forecasting methods for Autocorrelated Time Series, International Journal of Engineering Business Management, Vol.4, No.4, 2012.

[13] Kandananond. Forecasting Electricity Demand in Thailand with an Artificial Neural Network Approach, Energies, Vol.4, 1246-1257, 2011.

[14] N Slimani, I Slimani, N Sbiti and M Amghar. Traffic forecasting in Morocco Using Artificial Neural Networks, Elsevier, Procedia Computer Science, Vol. 151, 471-476, 2019.

[15] S. Raid, J. Mania, L.Bouchaou and Y. Najjar. Rainfall-runofff model using an artificial neural network approach, Mathematical and Computer modeling: An International Journal, 2004.

[16] A Apichottanakul, S Pathumnakul and Piewthongngam. Using an artificial neural network to forecast the market share if Thai Rice, IEEE-International Conference on Industrial Engineering and Engineering Management, 2009.

[17] S Pathumnakul, Piewthongngam and Apichottanakul. A neural network approach to the selection of feed mix in the feed industry, Elsevier, Computers and Electronics in Agriculture, Vol.68, No. 1, 2009.

[18] Chugai Biju R Mohan and G Ram Mohana Reddy. A Hybrid ARIMA-ANN Model for Resource usage Prediction, International Journal of Pure and Applied Mathematics, Vol. 119, No.12, 2018.

[19] Naveena, K., Subedar Singh, Santosha Rathod and Abhishek Singh. Hybrid ARIMA-ANN Modelling for Forecasting the Price of Robusta Coffee in India, International Journal of Current Microbiology and Applied Sciences, Vol.6, No.7, 2017.

[20] Bassi, S., Gomekar, A & Murthy, A.S.V. A learning algorithm for time series based on statistical features, Int J Adv Eng Sci Appl Math, Vol. 11, 230-235, 2019.

[21] Han, S. S., Azad, T. D., Suarez, P. A., and Ratliff, J. K. A machine learning approach for predictive models of adverse

events following spine surgery, The Spine Journal, Vol. 19, 2019.

[22] Yang, C., Delcher, C., Shenkman, E and Sanjay Ranka. Machine learning approaches for predicting high cost high need patient expenditures in health care, BioMed Eng OnLine 17, No.131, 2018.

[23] Takeshima T, Keino S, Aoki R, Matsui T, and Iwasaki K. Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data, Value in Health, Vol. 21, No. 2, 2018.

[24] Nti, I.K., Adekoya, A.F. & Weyori, B.A. A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data 7, No.20, 2020.

[25] Assa, H., Pouralizadeh, M., and Badamchizadeh, A. Sound Deposit Insurance Pricing Using a Machine Learning Approach, Risks, Vol. 7, No. 2, 2019.

[26] Pesantez-Narvaez, J., Guillen, M., and Alcaniz. M. Predicting Motor Insurance Claims using Telematics Data-XGBoost Versus Logistic Regression, Risks, Vol.7, No.2, 2019.

[27] Box, George E. P & Jenkins, Gwilym M. Time series analysis: forecasting and control. Holden-Day, San Francisco, 1970.

[28] G. Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model, Elsevier, Neurocomputing, Vol. 50, 159-175, 2003.

[29] P. Chujai, N Kerdprasop and K Kerdpraso. Time series analysis of household electric consumption with ARIMA and ARMA models, Proceedings of the International Multi Conference of Engineers and Computer scientists, Hong Kong, Vol. 1, 2013.

[30] X. Wang, and M. Meng. A Hybrid Neural Network and ARIMA Model for Energy Consumptions Forecasting, Journal of Computers, Vol.7, No.5, 1184-1190, 2012.