Volume 9, Issue 7, July – 2024
ISSN No:-2456-2165
International Journal of Innovative Science and Research Technology
https://doi.org/10.38124/ijisrt/IJISRT24JUL1269

An Analytical Approach to Predict Auto Insurance Claim using Machine Learning Techniques

Heena Kouser[1]
Student, Department of MCA
Jawaharlal Nehru National College of Engineering
Shivamogga, India

Hemanth Kumar[2]
Associate Professor, Department of MCA
Jawaharlal Nehru National College of Engineering
Shivamogga, India

Abstract:- Machine Learning business is regular in the insurance market to enhance the efficiency and predictive skills of the insurance industry linear regression as an initial and effective ml method is adopted in this work predicting automobile insurance claim is undertaken using these large datasets which provides the detailed driver characteristics vehicle characteristics and historical claim insights becomes possible to apply linear regression analyses to indicate and predict the likelihood and frequency of upcoming claims to insurers One of the primary motivations for linear regression is a very easy tool has it is pretty simple to use, easy to interpret and at the same time easy to scale it can benefit in managing and resolve the premium pricing and improvement of risk assessment along with the enhancement of the financial stability while applying linear regression the study explains how it can be utilized in auto insurance claims prediction the potential idea of using better ml models for more investigation and its pros and cons this model is mainly assessed in line with its predictive accuracy utilizing metrics reminiscent of mse and r-squared (R^2).

Keywords:- *Machine Learning Techniques; Prediction Model; Auto Insurance Claim.*

I. INTRODUCTION

Many companies are already in the state of losing a significant share of their earnings due to the increase in technical methods in ml and data analytics where insurance forms half of the worst-hit industries all among emerging behavior noticeable in the insurance domain is the growing application of ML method utilizes for strengthening the risk evaluation methodologies fraud analysis and estimations of the future claims concerning automobile insurance thus accuracy is essential for the insurance companies to predict auto insurance claims properly to ensure a sustainable financial position increase premium receivables and enhance customers satisfaction insurance claim prediction has generally been done mainly by using actuarial means and analytical techniques that are most of the time it was very hard to explain the extensive amount of data sets generated in the society however with the rise of ml entries and patterns that were previously unreachable can now can be utilized as an avenue through which some mining of big data can occur linear regressor is also a supervised ML techniques which is typical for predicting auto insurance claims a group of learning is called linear regression in which we learn a model which predicts a tar-get variable often referred to as the dependent variable on the basis of values of one or more input variables or the independent variables age of the driver type of car historical driving record etc. in this case the target variable would be the insurance claim amount the influence of these factors on the size and incidence of insurance claims are complete and also measurable with the help of linear regression which seeks to find the closest linear relationship of the data points.

II. LITERATURE SURVEY

N. Patel and M. Trivedi in [1] has developed a complete ML framework for predicting automobile insurance claims comparing various algorithms including linear regression decision trees and random forest. J. Liu et. al. in [2] has evaluated the usefulness of the amount of classifiers which includes decision trees, RF and SVM using a dataset. Findings showed that gradient boosting outperformed logistic regression which was the example with accuracy of 85%. Work in [3] by Q. Zhang et. al. included a detailed evaluation of several group strategies results organizing which included many core learners produced the greatest accuracy of 92% of the methods tested this work demonstrated the durability of methods based on ensemble learning in handling parts of inputs and enhancing prediction accuracy. Work by M. Patel et. al. in [4] has used AI, big data analysis methods including neural networks and k-nearest neighbour. The overall results of neural networks provided improved results with 87% accuracy. Work by L. Chen et. al. in [5] looked at how well xgboost and logistic regression strategies worked to forecast the incidence of accident claims with limited data training found that the suitable model for this is logistic regression due to its great predictive performance and interpretability. Overview of R. Agarwal, S. Gupta, and P. Verma in [6] has provided mixed approach models by integrating logistic regression with boosting techniques like adaboost and catboost techniques. Results the mixed model which shows how combined approaches may increase to forecast the accuracy and the combined approaches catboost and logistic regression techniques had an accuracy of 90%. Work by A Malekipirbazari et al. in [7] aimed to model and predict risks of motor insurance claims using Bayesian net-works. This yielded an 83% accuracy rate. S Atasoy and B Y Gokgoz in [8] produced techniques on an automobile insurance claim dataset comprising decision trees, random forests as well as neural networks with an aim of identifying most efficient algorithm approach among them all. Random

International Journal of Innovative Science and Research Technology
https://doi.org/10.38124/ijisrt/IJISRT24JUL1269

forest model achieved 94%. Authors in [9] lee kwon and choi sy proposed deep learning models to forecast insurance claims by integrating vehicle telematics data including GPS data along with driver actions the authors of the study employed recurrent neural networks RNN to detect temporal regularities from the information RNN models had an accuracy rate of 88%. The observation conducted by E. Albrecher et. al. in [10] discussed how extreme gradient boosting xgboost Could be utilized to forecast insurance claims more effectively. Using this approach, it was found that RF model achieved 94% accuracy while NN model had 82% percent which made it. A Survey by F. Silva et. al. in [11] implemented random forests, and SVM as ML algorithms to identify the fraudulent vehicle insurance claims. SVM model accuracy was 83%, while random forest was at 86%. G. Zhang et. al. in [12] investigated ensemble learning methods, such as stacking, boosting and bagging for predicting vehicle insurance claims. The stacking ensemble model was the most accurate, achieving 91% accuracy, which is nearly 3% higher than boosting techniques. Contribution by H. N. Nguyen et. al. in [13] proposed LSTM and CNN grids to predict auto insurance claims with chronological data. Mixing of CNN-LSTM model has a proof to encode the full Geographical and chronological patterns in this information with accuracy 90%. Work by I. W. Tsai et. al. in [14] introduced using different ML techniques such as ensemble methods, logistic regression and decision trees. The results confirmed the importance of ensemble learning for improving forecast quality 83%. Authors like T. Choudhury, A. Roy et. al. in [15] proposed usage of deep learning models, CNN and RNN, for predicting auto insurance claims with an accuracy of approximately 90%.

III. METHODOLOGY

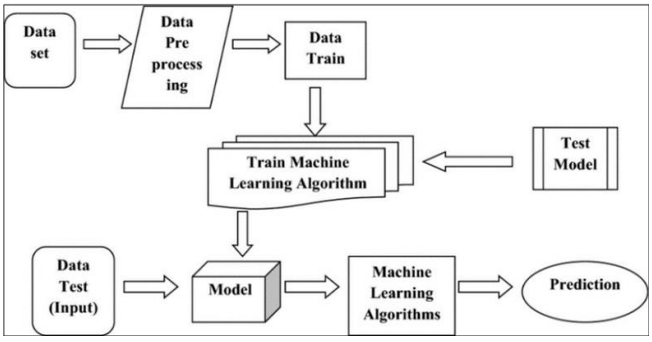

Fig 1 Machine Learning Zodel

The above Fig.1 illustrates ML Model is capable of recognizing patterns or making predictions on unseen datasets. ML is utilized in various fields for different purposes. Dataset is a collection of data for training and testing. Data pre-processing is used to clean the data for preparing a raw data into the format of ML algorithms. Train data will train the model using the given data as input. ML Algorithms are used to learn the patterns between input features and target variables. Test data will evaluate the performance and test the model using the provided data as output. Prediction result gives a trained ML Model when given a fresh input data.

A. Linear Regression Model

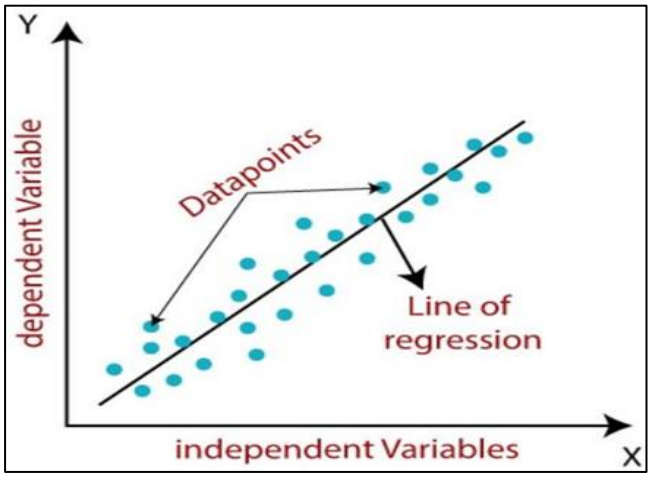

Fig 2 Linear Regression Model

Fig.2 represents linear regression, it is a statistical tool used to predict the cause-and-effect relationship between dependent variables and independent variables. However, this method is utilized to attain the estimation of the line in that model. This relationship as closely as possible. It works based on outcomes which dictate whether predicted results are proportionate to the numerical values fed into the model.

IV. PROPOSED MODEL

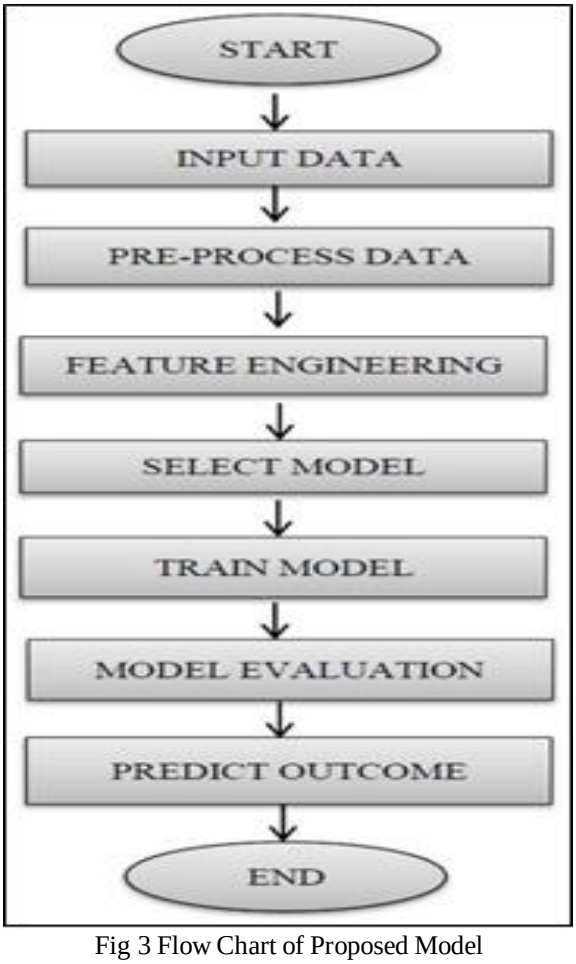

Fig 3 Flow Chart of Proposed Model

International Journal of Innovative Science and Research Technology
https://doi.org/10.38124/ijisrt/IJISRT24JUL1269

A. Data Description

Fig.3 shows the flow chart of proposed model. This work has accessed the data set from Kaggle site [16]. The model building is on training dataset to make it be a predictor of probabilities that claim has made or not. Table 1 below, contains some selected variables from the sets of details that is utilized for analysis and predicting the automobile insurance claim. While designing a model to predict car insurance claims is done by employing ML Techniques (MLT) using customers data.

Table 1 Description of the Dataset

Field Name	Description
policy_id	Unique identifier assigned to each insurance policy.
policy_tenure	Duration for which the policy has been active, usually in years.
age_of_car	The age of the insured car, typically measured in years.
age_of_policyholder	The age of the individual holding the insurance policy.
area_cluster	The age of the individual holding the insurance policy.
population_density	Measure of the population per unit area where the policyholder lives.
Fuel_Type	Type of fuel the insured car uses, such as petrol, diesel, or electric.
Model	Specific make and model of the insured vehicle.
airbags	The number of airbags installed in the insured car.
Gera_Box	Whether the car has an automatic or manual transmission system.
Ncap_rating	Safety rating of the car given by the New Car Assessment Program.
claim	Indicates whether a claim has been made under the insurance policy (binary outcome: 0 for no claim, 1 for claim).

B. Data Preprocessing

Cleanse the data information by handling missing values, outliers, and inconsistencies.

C. Feature Engineering

To handle the missing variables and the missing data analysis and it counts the Percentage of missing values.

D. Trainig and Testing

The information of data is divided into a train and test data in which the training is usually larger than test set.

E. Predict Outcome

In Fig.4 there is a distribution of Insurance Claims using Pie plot from the overall dataset with the interest of column chosen as target factor. Based on the density plot, 4-5-year-old cars are the most common for claim. The pie graph indicates that most cars do not claim, whereas only a few do. These together mean while some car ages are more common and higher risk of claiming.

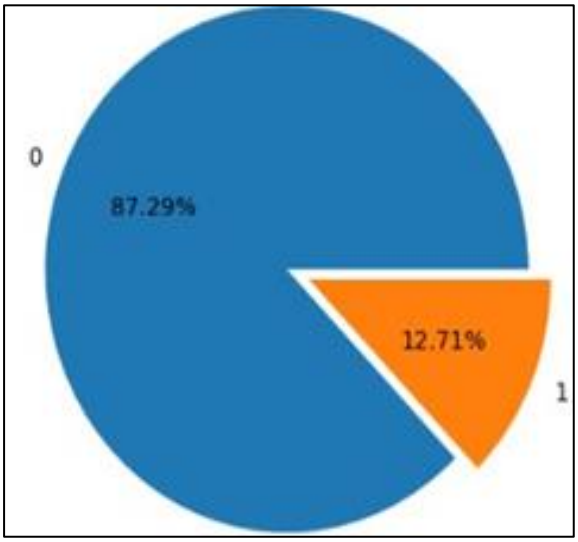

Fig 4 Pie Chart

In Fig.5 the distribution of Insurance claim using Univariate Analysis from complete dataset with the interest of column chosen as a target parameter. According to the univariate analysis, claimed percentage is more on old cars in recent 4-5 years.

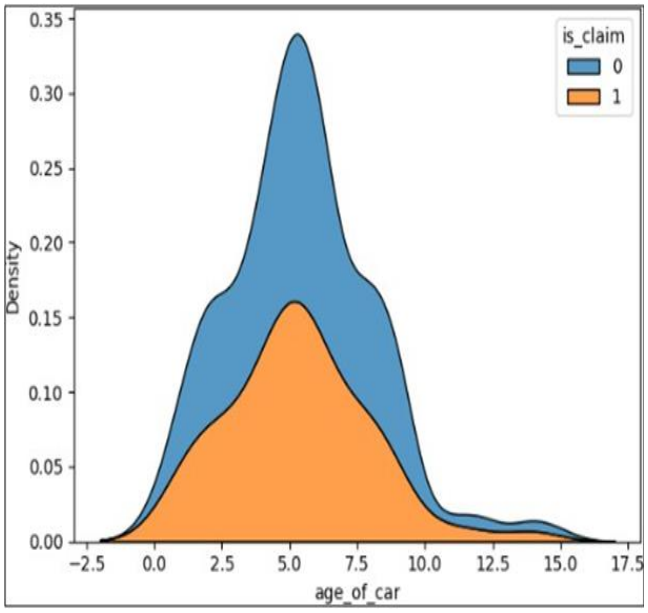

Fig 5 Univariate Analysis

V. MODEL EVALUATION

Model evaluation for linear regression has several metrics to measure the performance of the model. Here are the formulas. [17]

A. Mean Squared Error

MSE is the average of the squared errors. More sensitive to outliers than MAE.

$$MSE=1/n \sum (si - ti)\ 2 \qquad - (1)$$

Where si = observed value ti= predicted value

International Journal of Innovative Science and Research Technology
https://doi.org/10.38124/ijisrt/IJISRT24JUL1269

To find the MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all of those squared values and divide by the number of observations.

B. R-Squared (R^2)

R-squared is an essential statistical method in regression analysis that tells how well the independent variables explains the deviation in the dependent variable, the model illustrative power is measured by R^2 which compares the squared differences between real and anticipated values to the total deviation in the real values.

$$R^2 = 1 - \frac{\sum (si - ti)^2}{\sum (si - s\tilde{})^2} \quad \text{- (2)}$$

Where, $(si-ti)^2$ sum squared regression, it calculates how much the Real value deviates from the regression model's predictions. Smaller values reveals that the model's predictions are close to the actual data points. $(si- s\tilde{})^2$ total sum of squares It measures the total deviation in the dependent variable. It's a baseline measure indicating how much the data points deviate from the mean value.

VI. RESULTS

Confusion matrix and accuracy is used to assess the results of the classifier and it is also known as the contingency table is a specific table structure that shows the performance of a model.

Table 2 Confusion Matrix

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	TP	FN
Truth: Negative	FP	TN

Table 2 shows, True positives (TP): happen when the model correctly predict that the particular data is going to be positive. True negatives (TN): happen when the model points to a negative value and in reality, this is the case. False positives (FP): appear when the model assigns a positive prediction when the actual data point should have been negative. False negatives (FN): happen in the case when the model produces a wrong forecast and the actual, True Y, is less than the predicted Y.

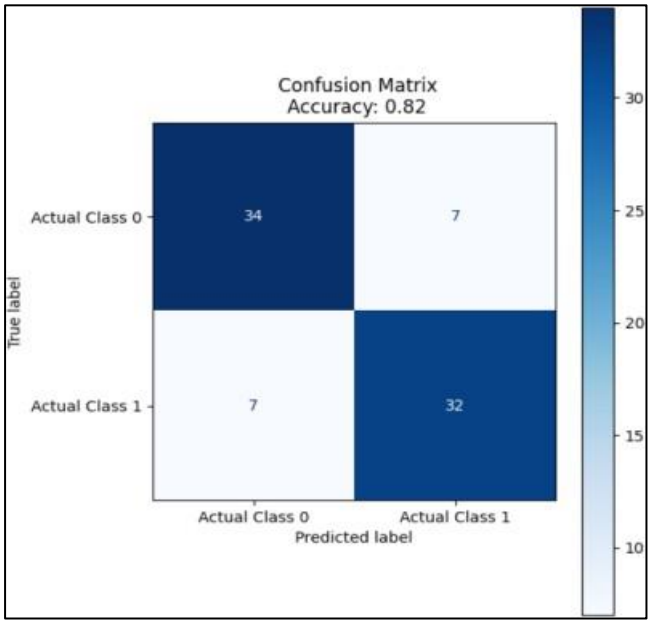

Fig 6 Confusion Matrix of Linear Regression

In Fig.6 Confusion matrix is generally linked with methods of classification when thresholding is implemented. Confusion matrix for dual classification can be produced from the forecasts of regression techniques offering data on the kind of classifications the model makes.

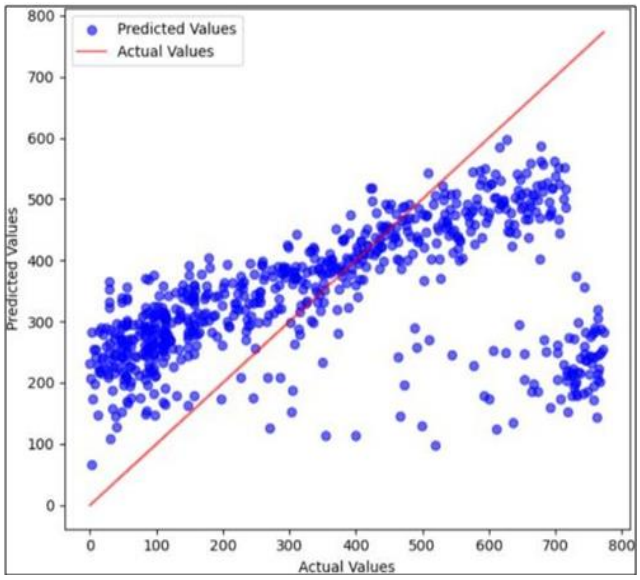

Fig 7 Linear Plot Actual vs. Predicted Values

Fig.7 shows the red line as actual values of dataset, and the blue dots explains predicted values Depending upon the given input by using the above dataset we can make more prediction by uploading live data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1269

VII. CONCLUSION

The linear regression approach was designed for forecasting automotive claim insurance has a moderate level of accuracy. The model finds critical characteristics that influence claim amounts, such as vehicle age, driver age, and policy Deductible. Furthermore, applying techniques such as decision trees or random forests to investigate non-linear correlations between characteristics and claim amounts may result in improved model performance. This work lays the groundwork for the creation of a more comprehensive claim prediction system, which will help to inform insurance pricing strategies, underwriting processes, and risk management activities.

REFERENCES

[1]. N. Patel and M. Trivedi, "Deep Learning for Auto Insurance Claim Prediction." Insurance: Mathematics and Economics, 2020, Vol 93, pp. 101-112, DOI 10.1016/j.insmatheco.2020.12.001.
[2]. J. Liu, H. Zhai, and Z. Wei, "Predicting Automobile Insurance Claims Using Gradient Boosting Machines", Journal of Data Science, 2017, Vol 15, No 1.
[3]. Q. Zhang, Y. Li, and X. Wang, "Hybrid Models for Predicting Auto Insurance Claims Using ML." Expert Systems with Applications, 2022, Vol. 188,115481, DOI: 10.1016/j.eswa .2022.115481.
[4]. M. Patel and N. Prajapati," Predictive Modelling for Auto Insurance Claims Using ML", International Journal of Computer Science and Information Security, 2019, Vol. 17, No 5
[5]. L. Chen, Y. Chen, and Z. Huang, "Deep Learning for Predicting Auto Insurance Claims", Expert Systems with Applications, 2020, Vol. 141.
[6]. R. Agarwal, S. Gupta, and P. Verma, "Hybrid ML Models for Auto Insurance Claims Prediction", Journal of Ambient Intelligence and Humanized Computing, 2021, Vol. 12, No. 6.
[7]. A. Malekipirbazari and V. Aksakalli, "Risk Prediction in Auto Insurance: A Bayesian Network Approach", European Journal of Operational Research, 2015, Vol. 242, No. 2.
[8]. Y. Gokgoz and S. Atasoy, "A Comparative Study of Different ML Techniques for Auto Insurance Claim Prediction", Procedia Computer Science, 2016, Vol 10.
[9]. D.Y. Lee, H. Kwon, and S. Y. Choi, "Predicting Auto Insurance Claims Using Deep Learning with Vehicle Telematics Data" IEEE Access, 2017, Vol. 5, DOI: 10.1109/ACCESS. 2017.268 50
[10]. E. Albrecher, H. Teugels and J. Beirlan," Insurance Claim Modelling and Prediction using Extreme Gradient Boosting", Astin Bulltin, 201 8, Vol. 48, No. 1, DOI: 10. 1017/ asb .2017. 36.
[11]. F. Silva, P. Cortez, and M. Santos, "Auto Insurance Fraud Detection with ML: A Case Study", Journal of Risk and Financial Management, 2019, Vol. 12, No. 2, DOI: 10.3390/ jrfm12020075.
[12]. G. Zhang, L. Jiang, and Y. Sun, "Utilizing Ensemble Learning Techniques for Auto Insurance Claim Prediction", Computers & Industrial Engineering, 2020, Vol. 14, DOI: 10. 1016/j.cie.2020 .106384
[13]. H. N. Nguyen, T. V. Nguyen, and Q. H. Nguyen, "Auto Insurance Claim Prediction Using Hybrid Deep Learning Models", Information Sciences, 2021, Vol. 546, DOI: 10.1016 /j.ins.2 020. 11.042
[14]. I. W. Tsai, C. H. Chen, and C. W. Huang, "Predictive Analytics for Auto Insurance Claims Using ML", Expert Systems with Applications, 2022, Vol. 186, DOI: 10.1016/j.eswa. 2021. 11 5946
[15]. T. Choudhury, A. Roy, and A. Paul, "Deep Learning for Auto Insurance Claim Prediction: A Case Study", IEEE Transactions on Neural Net- works and Learning Systems, 2020, Vol. 31, No. 9, DOI: 10.1109/ TNNLS. 2020.298057
[16]. https://www.kaggle.com/dataset
[17]. 8 ML Models Explained in 20 Minutes | Data Camp