International Journal of Multidisciplinary Research and Growth Evaluation.

Vehicle Insurance Purchase Prediction Using Machine Learning

Tamma Sreya [1*], Badugu Renuka [2], Sampathi Kalyan Chakravarthi [3], Nakkala Chaitanya Krishna [4], Mallela Narasimha Rao [5]

[1-4] Student, Department of Information Technology, Kallam Haranadhareddy Institute of Technology, (Autonomous), NH – 16, Chowdavaram, Guntur, Andhra Pradesh, India

[5] M. Tech, Assistant Professor, Department of Information Technology, Kallam Haranadhareddy Institute of Technology, (Autonomous), NH – 16, Chowdavaram, Guntur, Andhra Pradesh, India

* Corresponding Author: **Tamma Sreya**

Article Info

ISSN (online): 2582-7138
Volume: 06
Issue: 02
March-April 2025
Received: 03-02-2025
Accepted: 27-02-2025
Page No: 761-765

Abstracts

Aim: The research intends to Comparative Analysis of vehicle insurance prediction with Random Forest and logistic regression.

Materials and Methods: In this study, two groups Random Forest in comparison with Logistic Regression to improve Accuracy. To improve Accuracy. 100 dataset samples had been used for research study which contains 80% for training and remaining 20% for testing. For predicting vehicle insurance which were estimated by using a 10 N sample size for each,

Results: The Random Forest improves the data of accuracy with (93.605%). accuracy against (84.583%). for Logistic Regression. With a significant value of p= e of p=0.002 (p is statistically significant for prediction of vehicle insurance).

Conclusion: The Random Forest method for prediction of vehicle insurance significant improvement over Logistic Regression because of its higher accuracy.

DOI: https://doi.org/10.54660/.IJMRGE.2025.6.2.761-765

Keywords: Vehicle insurance, Logistic Regression, Random Forest, Machine Learning, Neural Network, Accuracy

Introduction

The effect of each characteristic on the general class was assessed, as well as the considerable incidence of fake high quality and false poor predictions using interpretable synthetic intelligence strategies. The results of this look at will assist threat analysts and professionals to assess the strengths and weaknesses of every version and increase empirically beneficial selection regulations for predicting future inputs. Kempa Severino (2021) In addition, in this text we have got compiled a comprehensive listing of showed frauds and analyzed them the use of synthetic intelligence (XAI) strategies, which explain the relative importance of the input variables in step with every model, and address false positives and giant motives. A fake nice. Preach Negative remarks. El Gayar (2019) They used Bayesian networks best to indicate claims or no claims, and used ML strategies to expect insurance fraud and insurance top rate amounts for specific clients based totally on non-public records; Three forms of ambush methods are used, Nandhini (2020), In this experience, using real records represents an extraordinary benefit over simulated records in each version evaluation and application to selection making. Hajek and Henriques (2020)

The IEEE Xplore Research Bank conducted the concept of faux user detection in 580 publications given 1,112 research articles submitted to the Google Scholar academic database. But it is unjust for a very good driving force to price more because of wherein they live; This creates a problem for the buyer that if the insurance charge is expanded, Kempa Severino(2021) However, deciding on the right strategies inclusive of characteristic choice strategies, function discrimination techniques, resuming algorithms, and ML classifiers can make contributions to the insurance policy's hard and unfavorable pleasant of loss prediction. (Math jack 2019) The fashions have been examined a couple of times and the average prediction consequences had been compared controlling for fake positives and false negatives. The outcomes show that ensemble-based methods (random forests and gradient boosting) and deep neural networks carry out higher, showing better common performance compared to

other classifiers, amongst which logistic regression is typically used (Hsu *et al*. (2020)**,** Road infrastructure has a big effect on avenue safety and desires in addition consideration. Based on road coincidence prediction, current research has started to broaden deep studying type algorithms. However, given the unsure nature of road injuries, the probabilistic estimation of avenue accidents on one-of-a-kind road segments remains a venture. In order to fill this knowledge hole, this look examined the actual music and gathered a twist of fate facts among 2013 and 2012. Then, in step with the spatial-temporal density ratio (Pts), the street segments are divided into three training, like low and medium. And the best danger of deportation. (Bo Wang,2020) Different varieties of fact matching algorithms have improved the overall performance of classifiers, classifiers predicting different varieties of composite overall performance metrics, and combined SMOTEENN and Random Forest algorithms have significantly stepped forward classification accuracy. In the destiny, the proposed avenue crash threat prediction approach may be tested greater in the assessment of road renovation and safety layout scenarios. (Chi Zhang 2021)

There are distinctive types of research to evaluate vehicle insurance. Previous algorithms were not able to detect imbalances in the dataset. Therefore, the main motive of car insurance is to misinform and lie to unsuspecting clients who need to be careful and try to defend themselves. The cause of the studies is to estimate automobile insurance, the usage of random forests and easy bases.

Materials and Methods

This is usually recommended research paintings finished at the Open-Source Lab**,** [[college name]]. Two businesses of category algorithms constituted the look at organization. Random wooded areas turned into protected in group 1, logistic regression became protected in group 2. In this study, each business has been compared with a random forest with logistic regression to enhance accuracy. To improve tactfully. A sample of 100 records was used for the studies observed, consisting of 80% for performance and 20% for checking out. Predictive vehicle coverage is anticipated by the usage of a pattern length of 10 N every, the attributes within the report are carefully designed so that the weather facts are applicable to the problem. A pattern of 100 facts was used for the studies, which include 80% for overall performance and 20% for checking out. The dataset became preprocessed by way of dividing it into education and testing, with 80% of the facts set aside for schooling and 20% for trying out. The dataset has zero values, making it appropriate for checking out diverse devices gaining knowledge of strategies.

The task became developed and evolved in Jupiter notebook the usage of Python programming. The Windows 10 operating gadget is the first machine mastering benchmark. The hardware layout includes 8 GB of RAM and an Intel Core i7 processor (8 GB). 64 bits are used for the hashed gadget. The code for the use of Python programming is implemented inside the language. The dataset becomes used to enforce the procedure to accurately code the procedure.

Random Forest

Random Forest is a system studying algorithms associated with the popular artwork of getting to know. M.L. It is based on the ensemble concept, including diversifications to correct the schooling problem and to boost the overall performance of the model.

As the name implies, "Random Forest is an algorithm that takes a couple of logistic regressions on various parameters of a given dataset and averages them to estimate the prediction accuracy of that dataset." Logistic regression - primarily based on decades, wherein the prediction of each tree is random and based on the range of votes, the tree. The variety of sets within the woodland ends in narrower and greater suitable to avoid the hassle. The photograph below suggests how the Random Forest algorithm works:

Algorithm

Step:1 The first step is to accumulate the facts you want to apply to educate the random woodland.

Step:2 Next, you want to define the trouble you want to resolve the usage of Random Forest. In this case it is far a binary trouble

Step:3 To examine a random leap model for performance, you need to divide the records into education and experiments**.**

Step:4 After the formation of the random woodland of samples, you can test its overall performance by looking at the set.

Step: 5 If the implementation of the random wooded area model is not first-class, you must alter the hyperparameters to get higher outcomes.

Step: 6 Once you are happy with the overall performance of the random bounce version, you can run it to predict new unseen data.

Logistic Regression

Logistic regression is one of the most powerful gears used inside the study of algorithms for both enterprise category and regression. It creates a flowchart-like tree structure wherein each internal node represents a characteristic certificate, each branch represents an output certificate, and every node includes a leaf (leaf node) label. It is built through iteratively dividing the education statistics into a fixed based totally on precise attributes till a final criterion along with the most peak or minimal variety of trees is reached. In schooling, the logistic regression set of rules selects the nice characteristic for the partition of data metrics consisting of entropy or Gini Impurity. The goal is to discover a hobby that will increase facts advantage or impurity discount after separation.

Logistic regression is a flower-leaf tree structure, in which every node represents an inner characteristic, branches constitute guidelines, and leaf nodes constitute the cease of the set of rules. The proposed system learning set of rules is a fashionable supervisor that can be implemented to both category and regression problems. It is a completely effective set of rules. And it can additionally be used to train random forests on a selection of training information sets, making random forests one of the most powerful algorithms in system mastering.

Algorithm

Step: 1 Identify the target variables and pick the relevant traits that affect the classification.

Step: 2 Standardize the information so that all attributes are identical**.**

Step:3 Set the initial values of the model parameters (weights and biases) to set or smaller random values**.**

Step:4 Define the sigmoid characteristic and the fee, which describes any actual quantity evaluated at a fee between 0 and 1**.**

Step: 5 Test the performance of the version using metric calculations such as precision, accuracy, recall and F1 rating.

Step:6 modify the above 6 parameters including getting to know price, regularization energy and variety of iterations to

enhance the version.
Step: 7 Use the learned logistic regression version to make predictions on the new data and identify them as real or fake jobs.

Statistical Analysis
Statistical analysis is performed using SPSS software program for random woodland statistical evaluation with logistic regression. IBM SPSS model 26 changed into used and a maximum of 10 iterations of each technique, consisting of IP cope with, geolocation, historical information, tool facts, person behavior, information and social media used, were analyzed as impartial variables and established variables. The vehicle insurance algorithms had been run on a 64-bit Windows 10 computer with 8GB of RAM, a secure Internet connection, and a Jupiter laptop loaded with Python 3.9. For precise organization details, use the organization ID. A five-pattern dataset is available for logistic regression and random wooded area

Results
Random forest and logistic regression have been tested on different times of Anaconda Navigator with a pattern size of 10.
Table 1 affords the exceptional information approximately the accuracy of the random forest.

Table 2 presents the anticipated accuracy of the logistic regression. Using those 10 statistical models, each rule was appropriately associated with the values used to evaluate the statistical kinds used.

Table 3 indicates the accuracy rankings of the Random Forest algorithm with logistic regression. Estimated values from random forests and logistic regression, using a mistake of 1.55589, 0.49201 and 1.88449,0.59593 respectively.

Table 4 compares the accuracy rating (93.605%) of Random Forest. Logistic regression (84.583%).

Figure 1 compares random woodland with logistic regression in phrases of average accuracy. Overall mean, popular error, and standard deviation for the random wooded area set of rules (93.605%). 1.55589, 0.49201 they stored. For logistic regression, the effects of the mean, well known deviation, and preferred errors of the primary are (84.583%),1.88449,0.59593respectively.

The random woodland algorithm has the mean, widespread error, fashionable deviation, and (93.605%) values of,1.55589, 0.49201 respectively. Car coverage consumer analysis evaluation sample size 10N, G value electricity 84.583% and statistically extensive fee z value of p = 0.02 (pelt; 0.05)1.88449,0. 59591.Logistic regression, suggest and comparative analysis amongst several random woodland category algorithms are all plotted. The category accuracy of one report is (93.605%). For logistic regression (84.583%), the random forest algorithm is greater correct.

Discussion
Today there are important methods in assessing traffic threats. As a method to pick out accident-susceptible avenue sections based on coincidence information, sufficient commentary time and pattern size are necessary to appropriately perceive accident-prone street sections. Such a technique has a certain put off and consequently it is tough to apprehend the plan and the beginnings of the operations of the roads. (Min Zhang 2020) Compared to conventional danger prediction methods, using selection trees and classifiers is more exploratory than adaptive. In this manner, higher danger elements can be taken into consideration, even capacity hazard factors whose mechanistic results are hidden. (Yiik Diew Wong 2022) Pricing gear typically performs in the context of generative linear modeling (GLM). With the rise of records analytics, our look specializes in device getting to know techniques to generate personal guidelines based totally on both the frequency and severity of requests. We manage the loss functions used within the algorithms to carefully incorporate the precise traits of facts warranty: enormously choppy compute data with many zeros and scatters, but long-lived records. Web page (Roel Henckerts 2020) We advise a higher technique for calculating insurance fraud detection financial savings. In addition to the saving effectiveness of 77 fraud detection methods the use of heats and the saving effectiveness of traditional statistical strategies primarily based on learning tools based totally on Romanian insurance strategies. The effects of this look at show that only a small percent of the to be had coverage fraud detection strategies may be used in a simply price-effective manner, and in fashionable, gadget gaining knowledge of detection methods have verified to be less cost-effective than traditional ones. Economic methods. Statistical machine. (Botond Benedek2021)

Tables and Figures

Table 1: The accuracy of random woodland values in the analysis of vehicle coverage (93.605%).

Iterations	Accuracy (%)
1	93.71
2	92.41
3	93.56
4	91.13
5	92.77
6	94.06
7	92.50
8	96.40
9	93.85
10	95.66
Accuracy	**93.60**

Table 2: Accuracy values of Logistic Regression in the Fraud detection is (84.5830%).

Iterations	Accuracy (%)
1	85.73
2	83.20
3	82.72
4	83.63
5	84.90
6	82.61
7	86.76
8	84.86
9	88.26
10	83.16
Accuracy	84.58

Table 3: Group Statistical Analysis of Random Forest and Logistic regression. Mean, Standard Deviation and Standard Error Mean are obtained for 10 samples. Random Forest has higher mean accuracy when compared to Logistic regression.

	Group	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	Random Forest	10	93.605	1.55589	0.49201
	Logistic regression	10	84.583	1.88449	0.59593

Table 4: Comparison of the Random Forest and Logistic regression with their accuracy

Classifier	Accuracy (%)
Random Forest	93.605
Logistic regression	84.583

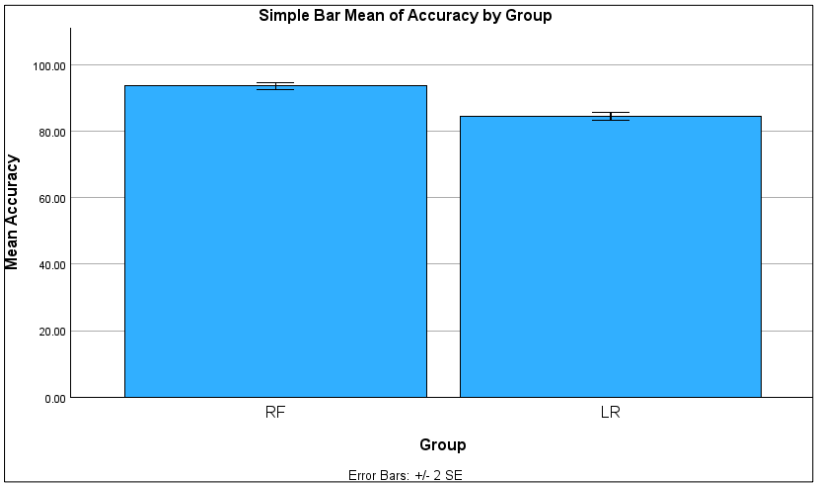

Fig 1: A comparison of random and unfastened regression based totally on common accuracy. The average accuracy of random containers is higher than that of logistic regression. It is slightly better than the usual logistic regression of random boxes. X-axis: random variety and logistic regression classifier and Y-axis: mean accuracy +/- 2 SD.

Conclusion

The essential purpose of these paintings is to assess how well Logistic Regression performs in the feature category using a random woodland category, a machine gaining knowledge of methods to the analysis of identification insurance and vice versa. The accuracy of the Random Forest classifier is 93.605%. The most correct method is Logistic Regression 84.583% Only the Random Forest classifier (93.605%) is anticipated to perform the Logistic Regression (84.583%). This changed into done so one can choose the dataset in this look at.

References

1. Doshi S. Traffic Sign Detection using Convolutional Neural Network. 2019.
2. de Deign J. Automatic Car Damage Recognition using Convolutional Neural Networks. 2018.
3. Manju M, More E. Research Paper on Intelligent Farming for Farmers using Control Systems in IoT. International Journal of Advanced Research in Computer Science. 2018;9(3):181.
4. Patil K, *et al*. Deep learning-based car damage classification. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE; 2017.
5. Harshani WAR, Vidanage K. Image processing-based severity and cost prediction of damages in the vehicle body: A computational intelligence approach. 2017 National Information Technology Conference (NITC). IEEE; 2017.
6. Kinsella G. Car damage monitoring system: Final project report analysis and design. Diss. Dublin: National College of Ireland; 2017.
7. Huang G, *et al*. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
8. He K, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
9. Jayawardena S. Image based automatic vehicle damage detection. 2013.
10. Hutter M, Gould S, Hartley R, Li H. Image Based Automatic Vehicle Damage Detection. 2013.
11. Ting HN, editor. 5th Kuala Lumpur International Conference on Biomedical Engineering 2011: BIOMED 2011, 20-23 June 2011, Kuala Lumpur, Malaysia. Vol. 35. Springer Science & Business Media; 2011.
12. Sharma N, Banga VK. Drowsiness warning system using artificial intelligence. World Academy of Science, Engineering and Technology. 2010;4(7):1771–3.
13. Li Y, Dorai C. Applying Image Analysis to Auto Insurance Triage: A Novel Application. 2007 IEEE 9th Workshop on Multimedia Signal Processing. IEEE; 2007.
14. Sun Z, Bebis G, Miller R. On-road vehicle detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;28(5):694–711.
15. Stromile P. Traffic Sign Detection using Convolutional Neural Network. Segmentation.
16. Harai S, Khatri SK, Singh G. Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique. Nov 2019. Available from: https://ieeexplore.ieee.org/document/9036162.
17. Itri B, Mohamed Y, Mohammed Q, Omar B.

Performance comparative study of machine learning algorithms for automobile insurance fraud detection. 30 Oct 2019. Available from: https://ieeexplore.ieee.org/document/8942277.
18. Muranda C, Ali A, Shongwe T. Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method. 16 Oct 2020. Available from: https://ieeexplore.ieee.org/document/9259322.