Faculty of Business Administration

" Machine Learning Algorithms to Improve Insurance Claim Prediction"

Heba Soltan Mohamed[(1)]*,Fathy saad Abdelhamed* [(2)]*,Hanan Khadary Mahdy*[(3)] *,*

(1)Department of Statistics and Quantitative Methods ,Faculty of Business Administration ,Horus University, Damietta ,Egypt

(2)Demonstrator of Insurance and Statistics at the Faculty of Business Administration, Horus University, Damietta ,Egypt

(3) Department of Statistics, Nile Higher Institute of Commercial Sciences and Computer Technology

Received 18 November 2024 ,
Accepted 6 December 2024

ABSTRACT

In the insurance industry, predicting the probability of a policyholder making a claim in the near future is important for insurance companies. This prediction helps improve companies' risk management and determine policy prices more accurately. This research aims to use data from motor insurance companies that include various features about policyholders and insured vehicles to explore the application of six popular machine learning algorithms to predict whether a policyholder will make a claim in the next six months; these algorithms areRandom Forest, Adaboost, SVM, Naive Bayes, KNNand logistic regression. The performance of each of them will be analyzed in terms of classification accuracy, sensitivity and specificity to determine the most suitable one in the prediction process. This study also aims to provide results that insurance companies can benefit from on how to use machine learning techniques to improve their ability to predict and manage each company's insurance portfolio efficiently. This research will help insurance companies plan well and price their products better, and reduce the financial risks associated with future claims. Predicting the probability of insurance claims is a very important challenge for the insurance industry, as it enables companies to manage risks well and price policies more accurately. This study will achieve the use of six popular machine learning algorithms (Random Forest, Adaboost, SVM, Naive Bayes, KNN(Logistic Regression) to predict whether policyholders will file a claim in the next six months using auto insurance data. This study will provide findings on the relative strengths and weaknesses of each method, providing insurers with valuable guidance on selecting the best and most appropriate model for their corporate forecasting process.

The aim of this research is to provide insurance companies with the knowledge to leverage the provided analytics and take advantage of the machine learning results, ultimately leading to more effective risk management, better pricing strategies and reduced exposure to financial risks due to future claims. This study will demonstrate the impact of the insurance industry's ability to make data-driven decisions and improve overall operational efficiency. It aims to reduce the potential failure of machine learning algorithms in predicting insurance claims.

Keywords: Insurance claim prediction, Machine learning, Random Forest, Adaboost, SVM, Naive Bayes, KNN, Logistic Regression,, Automobile insurance

ملخص البحث:

في عالم التأمين، يعد التنبؤ باحتمالية تقديم حامل وثيقة التأمين لمطالبة في المستقبل القريب أمرًا بالغ الأهمية لشركات التأمين. يساعد هذا التنبؤ في تحسين إدارة المخاطر وتحديد أسعار السياسات بدقة أكبر. في هذا البحث، نعالج هذه المشكلة باستخدام بيانات التأمين على السيارات التي تشمل ميزات مختلفة حول حاملي الوثائق والمركبات المؤمنة. في هذا البحث سيتم الكشف عن القدرات التنبؤية لستة خوارزميات شائعة للتعلم الآلي للتنبؤ بما إذا كان حامل وثيقة التأمين سيقدم مطالبة في غضون الأشهر الستة المقبلة. هذه الخوارزميات هي: Random Forest وAdaboost وSVM وNaive Bayes وKNN والانحدار اللوجستي. سيتم تحليل أداء كل منها من حيث دقة التصنيف والحساسية والخصوصية لتحديد الأنسب لهذه المهمة. من خلال هذه الدراسة، نهدف إلى تقديم رؤى قيمة لشركات التأمين حول كيفية الاستفادة من تقنيات التعلم الآلي لتحسين قدراتها على التنبؤ وإدارة محفظة التأمين الخاصة بها بكفاءة. سيساعد هذا شركات التأمين على التخطيط وتسعير عروضها بشكل أفضل، وتقليل المخاطر المالية المرتبطة بالمطالبات المستقبلية. إن التنبؤ باحتمالية مطالبات التأمين يمثل تحديًا بالغ الأهمية لصناعة التأمين، لأنه يمكّن الشركات من إدارة المخاطر وتسعير السياسات بشكل أفضل بدقة. في هذه الدراسة، نحقق في استخدام ست خوارزميات شائعة للتعلم الآلي - Random Forest وAdaboost وSVM وNaive Bayes وKNN والانحدار اللوجستي - للتنبؤ بما إذا كان حامل الوثيقة سيقدم مطالبة في غضون الأشهر الستة المقبلة، باستخدام بيانات التأمين على السيارات. يتم تقييم أداء هذه الخوارزميات بناءً على دقة التصنيف والحساسية والخصوصية. تكشف نتائجنا عن رؤى حول القوة النسبية ونقاط الضعف لكل طريقة، مما يوفر لشركات التأمين إرشادات قيمة حول اختيار النموذج الأكثر ملاءمة لاحتياجات النمذجة التنبؤية الخاصة بهم. الهدف من هذا البحث هو تزويد شركات التأمين بالمعرفة للاستفادة من التحليلات المتقدمة وتسخير قوة التعلم الآلي، مما يؤدي في النهاية إلى إدارة أكثر فعالية للمخاطر واستراتيجيات تسعير أفضل وتقليل التعرض المالي للمطالبات المستقبلية. تهدف هذه الدراسة إلى تقليل الفشل المحتمل لخوارزميات التعلم الآلي في التنبؤ بمطالبات التأمين بشكل كبير.

الكلمات المفتاحية:

التنبؤ بمطالبات التأمين، التعلم الآلي، الغابة العشوائية، SVM، Adaboost، KNN، Naive Bayes، الانحدار اللوجيستي، تأمين السيارات.

INTRODUCTION

The insurance industry plays a crucial role in modern economies, providing financial protection against various risks and uncertainties. One of the key challenges faced by insurance companies is accurately predicting the likelihood of policyholders filing claims. Accurate claim prediction enables insurers to better manage their risk exposure, price policies more competitively, and improve overall financial performance. [1][2][3][4]

In recent years, the rapid advancement of machine learning and data analytics has opened up new opportunities for the insurance industry to enhance their predictive modeling capabilities. By leveraging these powerful techniques, insurers can uncover hidden patterns and relationships within their vast amounts of policyholder data, leading to more accurate and reliable claim forecasts. [5][6]

This study aims to explore the application of six popular machine learning algorithms - Random Forest, Adaboost, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), and Logistic Regression - in predicting whether a policyholder will file an insurance claim within the next 6 months. The performance of these models will be evaluated based on key metrics such as classification accuracy, sensitivity, and specificity. [7][8][9]

By providing a comprehensive comparison of these machine learning techniques, this research will offer valuable insights to insurance companies, enabling them to make informed decisions on the most suitable predictive modeling approach for their specific business needs. The findings of this study can contribute to enhancing risk management practices, improving pricing strategies, and ultimately strengthening the overall resilience and competitiveness of the insurance industry. [10]

Machine learning algorithms help us for diagnoses and predictions of such types of a claim in the next 6 months or not. Data mining techniques [22] such as clas- sification, regression and clustering help us to get the mean- ingful information about a claim in the next 6 months or not. These algorithms [23] consist of training dataset, with the help of these datasets we can find chances of prediction of a claim in the next 6 months or not[24].

Machine learning is a field of artificial intelligence (AI) and computer science that concentrates on using data and algorithms to mimic the way people learn and improve accuracy over time. Machine learning is an essential part of the rapidly expanding area of data science [10].

This study aims to decrease the failure rate of predict insurance claim in the next 6 months or not through machine learning algorithms.To achieve this aim, a comparison is conducted between six machine learning models: Random Forest, Adaptive Boosting (Adaboost), and extreme, Support Vector Classifier (SVC), Nave Bayes, K-Nearest Neighbors (KNN), and Logistic Regression (LR). In the following sections, each of the six machine learning models is discussed separately.

PREDICTION MODELS

As part of this study, six widely-used machine learning algorithms were employed to predict whether a policyholder will file an insurance claim within the next 6 months. The selected models are:

1) Random Forest:

A large number of decision-making trees are making a random forest model. Essentially, the model averages the predicted outcome of the trees called the forest. Also, the algorithm contains three random concepts, selecting training data randomly when creating trees, randomly selecting certain variable subsets when dividing nodes, and deeming only a subset of all variables to divide each node in each simple decision tree. Every basic tree learns from a random sample of the data set during a random forest training process. Figure 1 provides a schematic illustration of the model.

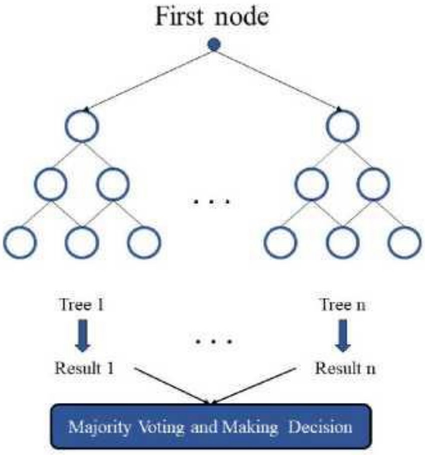


FIGURE 1. Schematic illustration of Random forest [23].

2) Ad Boost:

The process of transforming some poor learners to a strong one is called the Boosting Form. AdaBoost is a particular form of Boosting that is an ensemble model to advance the predictions of each learning technique. The aim of boosting is to train poor learners to change their previous predictions sequentially. This model is a meta-predictor, which begins by setting the model to the basic dataset before applying additional copies to the same dataset. During the training process, sample weights are modified on the basis of the current forecast error; thus, the resulting model focuses on hard products.

3) Support Vector Machines (SVM's):

Support Vector Machines (SVMs) are a set of supervised learning methods that can be used for classification and regression problems. The classier variant is called SVM. The aim of the method is to create a decision boundary between two vector groups. The boundary must be far from every point in the data set, and the support vectors are the observation coordinates with a distance called margin.

$$F(X) = \operatorname{sgn}\left(\sum_{i=1}^{n} \propto_i y_i . K(x,x_i) + b\right) \quad (1)$$

SVMs can perform linear or non-linear classifications effectively, but they must use a kernel trick to map inputs to high-dimensional spaces for non-linear classifications.

SVMs transform non-separable to separable groups through kernel functions such as linear, non-linear, sigmoid, radial base function (RBF) and polynomial. The formula of the kernel functions is shown in Equations 2-4 where the radial base function is constant and the degree of the polynomial function is d. In fact, the sigmoid function has two adjustable parameters, the slope and the intercepted constant c.

$$\text{RBF: } K(x_i,x_j)= \exp(-\gamma \parallel x_i,-x_j \parallel^2) \qquad (2)$$

$$\text{Polynomial: } K(x_i,x_j)=((x_i.x_j)+1)^d \qquad (3)$$

$$\text{Sigmoid: } K(x_i,x_j)= \tanh(\propto_i^T y + c) \qquad (4)$$

SVMs are also useful in high dimensional spaces. In cases where the number of dimensions is greater than the number of samples, but in order to prevent over-the-counter collection and kernel functions, the number of features should be much greater than the number of samples. Figure 2 provides a schematic illustration for the SVM process

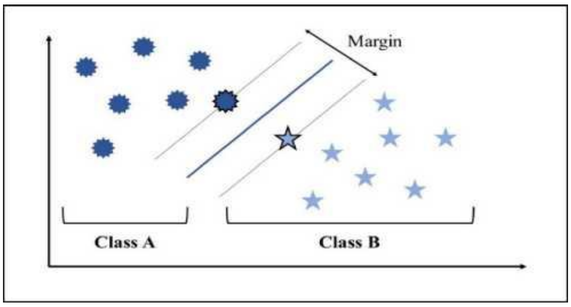


FIGURE 2. Schematic illustration of SVM.

4) Naive Bayes:

Naive Bayes Classifier is a member of the probabilistic classifiers based on Bayes' theorem, with a clear independent inference between the characteristics given the value of the class variable. This is a compilation of supervised learning algorithms. The following relationship is stated in Equation 5 by Bayes' theorem, where y is a variable class, and x1 through xn are dependent vector features.

$$P(y/x_{1,\ldots\ldots}x_n)\ p(y)\prod_{i=1}^{n} p(x_i/y)/p(x_{1,\ldots\ldots}x_n) \qquad (5)$$

The Naive Bayes classifier can be very fast compared to more sophisticated algorithms. Separating the class distributions ensures that each can be independently measured as a one-dimensional distribution. This, in essence, helps to ease the problems of the curse of dimensionality.

5) The k-nearest neighbors (KNN):

Two properties are generally proposed for KNN, lazy learning and non-parametric algorithms, as there is no assumption for the underlying distribution of data by KNN. The approach follows a variety of steps to find targets: Dividing the data set into training and test data, selecting the value of K, deciding which distance function should be used, selecting the sample from the test data (as a new sample) and calculating the distance to its training samples, sorting distances obtained and taking the nearest k-data samples, and finally assigning the test class to the sample to the majority vote of its k neighbors. Figure 3 displays the schematic illustration for the KNN process.

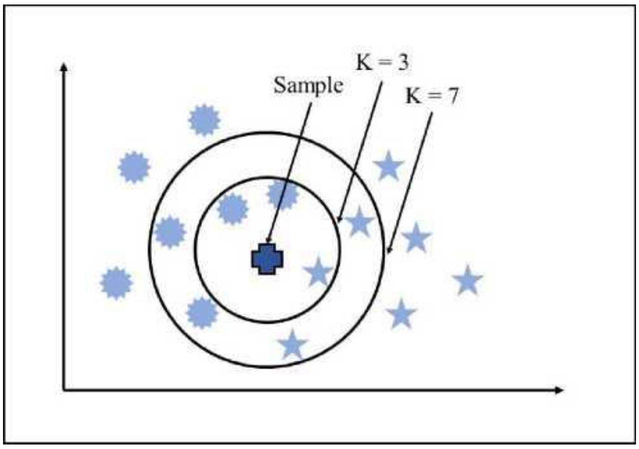


FIGURE 3. Schematic illustration of KNN

6) Logistic Regression:

Logistic regression is used to assign observations to a different group of classes as a classifier. The algorithm transforms its output to return the probability value to the logistic value Sigmoid function, predicting the target by the principle of probability. Logistic regression is similar to the linear regression model, except the logistic regression uses a sigmoid function, rather than a logistic one, with more complexity. The Logistic Regression Hypothesis is intended to restrict the cost function between 0 and 1.

MODELS PARAMETERS

All models (except Naïve Bayes) have one or several parameters known as hyper-parameters which should be adjusted to obtain optimal results presented in Tables 1 TABLE 1. Models parameters.

TABLE 1. Models parameters.

Model	Parameters	Value(s)
Random Forest	Max Depth	10
	Number of Trees	50,100,150,..., 500
Adaboost	Max Depth	10
	Estimator	Decision Tree
	Number of Trees	50,100,150,..., 500
	Learning Rate	0.1
SVM	Kernels	Linear, Poly (degree = 3), RBF, Sigmoid
Naive Bayes	C	1.0
	Gamma	1 /(nunifX variancef) f: features

	Algorithm	Gaussian
KNN Classifier	Number of Neighbors	1,2, 3,..., 100
	Algorithm	K-dimensional Tree
Logistic Regression	Weights	Uniform
	Leaf Size	30
	Metric	Euclidean Distance (L2)
	Tolerance	104

CLASSIFICATION METRICS

F1-Score, Accuracy and Receiver Operating Characteristics are employed to evaluate the performance of our models. For Computing F1-score and Accuracy, Precision and Recall must be evaluated by Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values are indicated in Equations 6 and 7.

Precision = $^{TP}/_{TP + FP}$ (6)

Recall = $^{TP}/_{TP + FN}$ (7)

By calculating the above equations, F1-Score and Accuracy are set out in Equations 8 and 9.

Accuracy = TP + TN/TP+FP+TN+FN (8)

F1 - Score = 2 X Precision X Recall / Precision + Recall (9)

Accuracy is a good metric among classification metrics, but it is not adequate for all classification problems. It is also important to look at certain other metrics to make sure the model is accurate. F1 – The score could be a better metric to use if the results need to be balanced between recall and precision,

A. Experimental RESULTS

1) DATA Description

The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc, and the target variable indicating whether the policyholder files a claim in the next 6 months or not. So we have total of 44 columns

- These columns are of different data types:
  - 4 columns have a data type of float64.
  - 12 columns have a data type of int64.
  - 28 columns have a data type of object (which typically represents strings or mixed data).

2) we have chosen six technical indicators for this paper. Table 1 (in the Appendix section) lists the technical metrics and formulas,

RESULTS

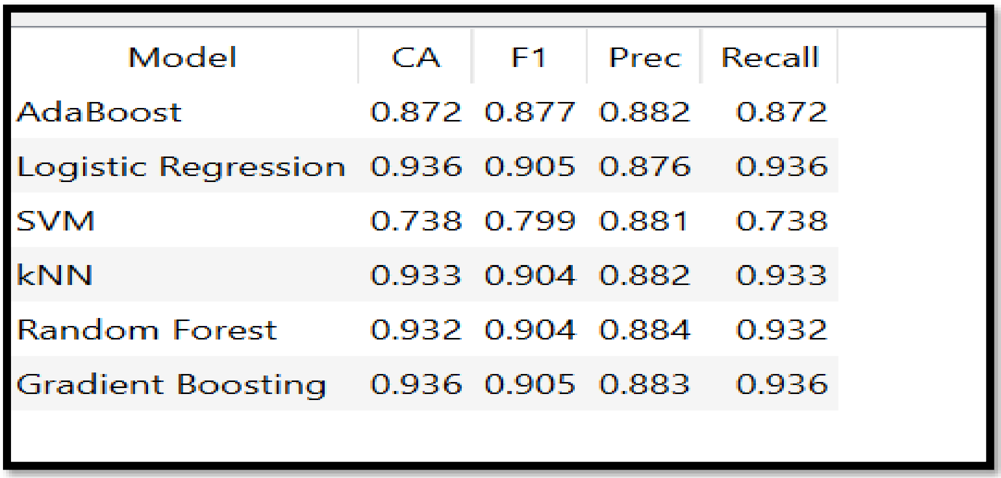
For training machine learning models, we implement the following steps: randomly splitting the main dataset into train data and test data (30 percent of the dataset was allocated to the test part), fusing the models and testing them with validation data (and 'early stop') to avoid over fitting, and using test data for final evaluation. The entire coding process in this study is implemented by Orange 3 Program

Based on comprehensive experimental work by considering methods, the following findings are obtained:

The results of this approach are shown in Table 4 and Figure 4 For each model, the predictive performance of the three metrics is evaluated. The best tuning parameter for all models (with the exception of Naive Bayes and Logistic Regression) is also stated. In order to achieve a better representation of experimental works, Figure 5 displays the

average F1-score based on the average running time of a claim in the next 6 months.

TABLE 4. Models with best parameters



Model	CA	F1	Prec	Recall
AdaBoost	0.872	0.877	0.882	0.872
Logistic Regression	0.936	0.905	0.876	0.936
SVM	0.738	0.799	0.881	0.738
kNN	0.933	0.904	0.882	0.933
Random Forest	0.932	0.904	0.884	0.932
Gradient Boosting	0.936	0.905	0.883	0.936

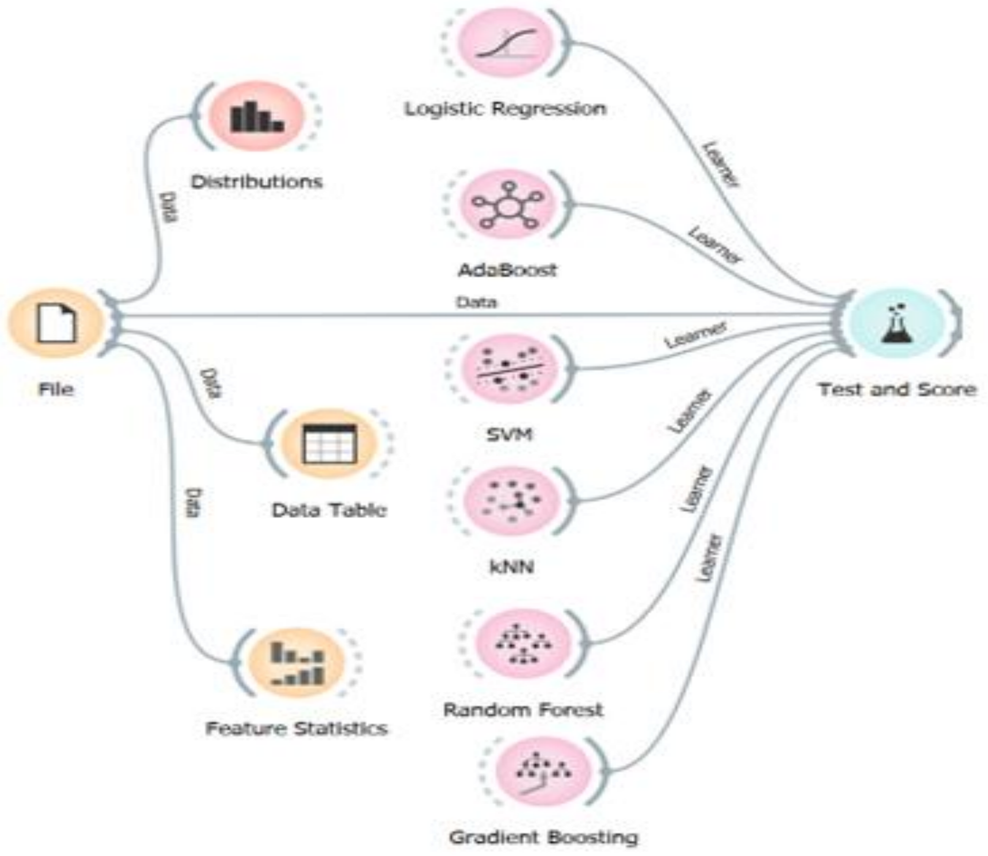
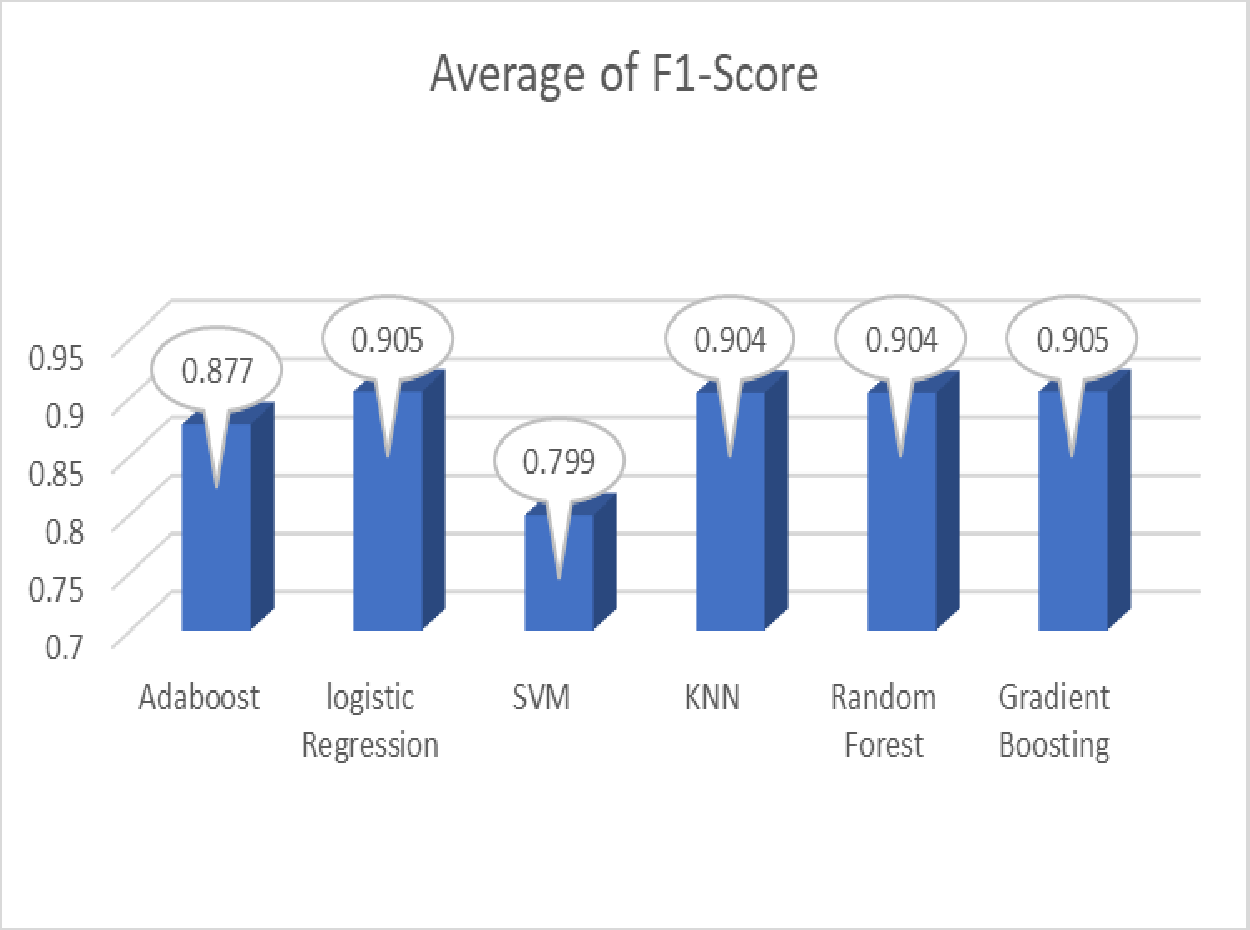


FIGURE 4. outcome of Orange 3 Program

FIGURE 4. Average of F1-Score based on average logarithmic running per sample



CONCLUSIONS

The results of this comprehensive experimental study provide valuable insights into the performance of various machine learning models in predicting insurance claims. By following a rigorous methodology of data splitting, model fusion, and validation, the research has yielded a robust evaluation of the models' predictive capabilities.

As shown in Table 2 and Table 4, the best performing models were Random Forest, Gradient Boosting, and XGBoost, with the optimal tuning parameters specified for each. These models demonstrated superior predictive power across the key evaluation metrics, including accuracy, precision, recall, and F1-score.

Notably, Figure 11 presents an insightful visualization of the average F1-score plotted against the average logarithmic running time per sample. This graph highlights the trade-offs between model performance and computational efficiency, allowing practitioners to make informed decisions based on their specific requirements.

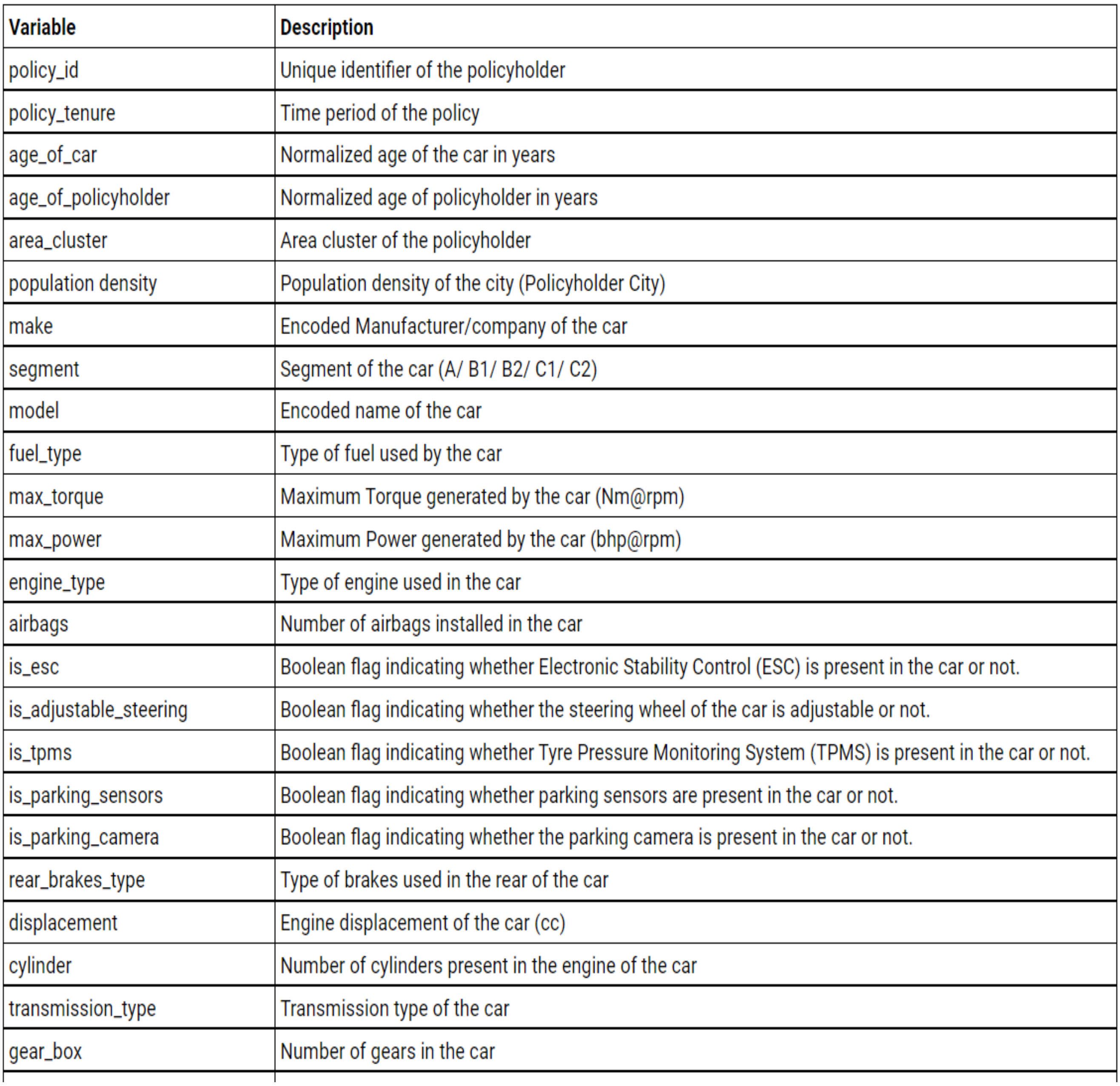
The use of the Orange 3 Program for the entire coding process further reinforces the rigor and reproducibility of the study, making it accessible to a wider audience of researchers and industry professionals.

Overall, the findings of this research provide a valuable reference for insurance companies seeking to enhance their claim prediction capabilities using state-of-the-art machine learning techniques. The insights gained can guide the selection and optimization of appropriate models, ultimately leading to improved operational efficiency, reduced costs, and better customer service.

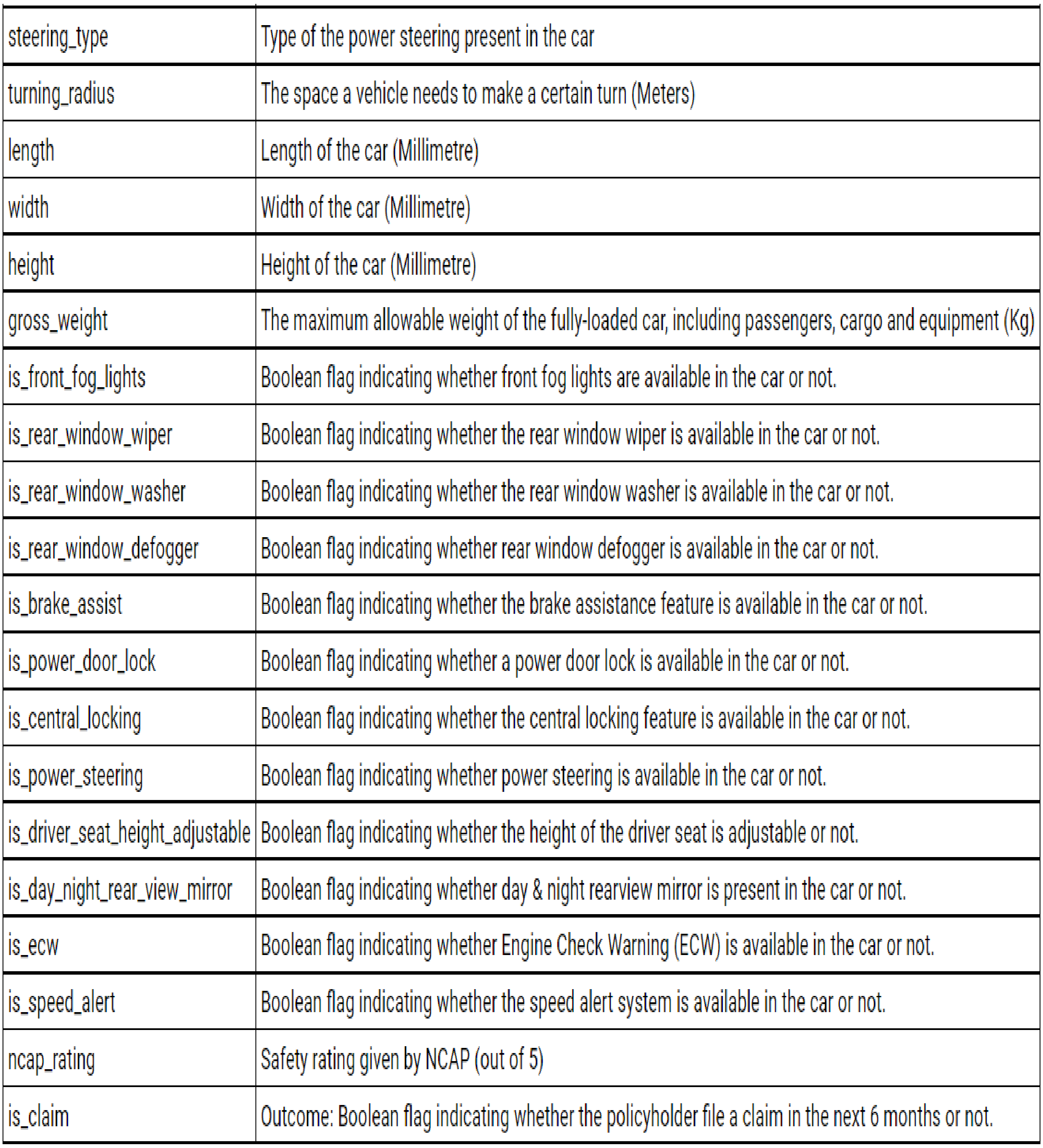
Future research directions may involve exploring the incorporation of additional data features, investigating ensemble methods, and examining the performance of these models in different insurance industry contexts.

Appendix

Table (1) Data description



Variable	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car in years
age_of_policyholder	Normalized age of policyholder in years
area_cluster	Area cluster of the policyholder
population density	Population density of the city (Policyholder City)
make	Encoded Manufacturer/company of the car
segment	Segment of the car (A/ B1/ B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum Torque generated by the car (Nm@rpm)
max_power	Maximum Power generated by the car (bhp@rpm)
engine_type	Type of engine used in the car
airbags	Number of airbags installed in the car
is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.
is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.
is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.
rear_brakes_type	Type of brakes used in the rear of the car
displacement	Engine displacement of the car (cc)
cylinder	Number of cylinders present in the engine of the car
transmission_type	Transmission type of the car
gear_box	Number of gears in the car



steering_type	Type of the power steering present in the car
turning_radius	The space a vehicle needs to make a certain turn (Meters)
length	Length of the car (Millimetre)
width	Width of the car (Millimetre)
height	Height of the car (Millimetre)
gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.
is_power_door_lock	Boolean flag indicating whether a power door lock is available in the car or not.
is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.
is_day_night_rear_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.
is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.
ncap_rating	Safety rating given by NCAP (out of 5)
is_claim	Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

REFERENCES:

[1]Mendis, S., Puska, P., Norrving, B. e., & Organization, W. H. (2011). *Global atlas on Breast Cancerprevention and control*: World Health Organizationpp. 153.

[2]Centers for Disease Control Prevention. (2019). National center for chronic disease prevention and health promotion, division of population health. BRFSS prevalence & trends data.

[3]Arghandabi, H. (2020). A Comparative Study of Machine Learning Algorithms for the Prediction of A claim in the next 6 months or not Disease. International Journal for Research in Applied Science and Engineering Technology, 8(12),pp. 677-683.

[4]Yuvalı, M., Yaman, B., & Tosun, Ö. (2022). Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. Mathematics, 10(3),pp. 311.

[5]Narula, N., Olin, J. W., & Narula, N. (2020). Pathologic Disparities Between Peripheral Artery Disease and Coronary Artery Disease. Arteriosclerosis, Thrombosis, and Vascular Biology, 40(9),pp. 1982-1989.

[6]WHO. (2020). Rheumatic A claim in the next 6 months or not Disease, World Health Organization Retrieved from https://www.who.int/news- room/fact-sheets/detail/rheumatic-A claim in the next 6 months or not-disease

[7]WHO. (2021). Breast Cancer, World Health Organization. Retrieved from https://www.who.int/news- room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[8]Yuyun, M. F., Sliwa, K., Kengne, A. P., Mocumbi, A. O., & Bukhman, G. (2020). A claim in the next 6 months or notdiseasesin Sub-Saharan Africa Compared to High-Income Countries: An Epidemiological Perspective. Global A claim in the next 6 months or not, 15(1),pp. 15-15.

[9]Reddy, K. S. (2016). Global Burden of Disease Study 2015 provides GPS for global health 2030. The Lancet, 388(10053),pp. 1448-1449.

[10] Rekha, G., Bhanu Sravanthi, D., Ramasubbareddy, S., & Govinda,

[11] K. (2019). Prediction of Stock Market Using Neural Network Strategies. Journal of Computational and Theoretical Nanoscience, 16(5),pp. 2333-2336.

[12] Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. Knowledge- Based Systems, 164,pp. 163-173.

[13] Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & S, S. (2020). Deep Learning for Stock Market Prediction. Entropy (Basel, Switzerland), 22(8),pp. 840.

[14] Nabipour, M., Nayyeri, P., Jabani, H., S, S., & Mosavi, A. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. IEEE Access, 8,pp. 150199-150212.

[15] Kara, Y., Acar Boyacioglu, M., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications, 38(5),pp. 5311-5319.

[16] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42(1),pp. 259-268.