ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 10, October 2021

COMPARISON OF FEATURE SELECTION AND CLASSIFICATION ALGORITHMS IN INSURANCE FRAUDULENT DATASET

M. Sathya
Research Scholar,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: msathya46@gmail.com

Dr. B. Balakumar
Assistant Professor,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: balakumarcite@msuniv.ac.in

Abstract — In terms of large amounts of data, the insurance business is rapidly developing. Fraudulent claims are the most serious problem in the insurance industry. Fraud is defined as a deceptive or unlawful act intended to acquire financial or personal benefit. The previous approach will no longer function as the number of data grows, and identifying fraudulent claims will become a time-consuming task. This paper compares various machine learning techniques for detecting fraud claims in the Automobile Insurance Claims datasets. In this study, a large dataset of automotive insurance claims is employed, and three feature selection techniques are applied to the dataset, with the classification algorithms using the features to identify fraud claims. The Feature Selection algorithms used in this paper are Tree-Based Feature Selection Algorithm, L1-Based Feature Selection Algorithm and Univariate Feature Selection Algorithm and the classification algorithms are Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbor (KNN). Performance measurements such as accuracy, precision, and recall are used to compare these algorithms. The suggested model demonstrates that Random Forest performs well in terms of accuracy and precision, whereas Decision Tree performs best in terms of retention.

Keywords — Automobile Insurance; Machine Learning; Classification Algorithms; Algorithm Comparison

I. INTRODUCTION

Insurance is a contract in which a person pays premiums to an insurance firm in exchange for protection against losses. There are around 1000 insurance companies in the world, and they collect nearly ten thousand crores rupees per year. Insurance fraud occurs when a person submits a fake insurance claim in order to gain from it. The entire cost is anticipated to be in excess of 400 crore rupees. As a result, predicting insurance fraud claims is a difficult task.

According to a survey conducted by India Forensic Research, the Insurance Sector in India loses roughly 30,401 Crore of rupees per year due to frauds; in other words, frauds account for 8.5 percent of the sector's revenue. The Insurance Sector is divided into two parts: Life Insurance and General Insurance. Automobile/Vehicle Insurance is the most profitable and largest segment of General Insurance. In 2009, Auto Insurance had a share of Rs.915 crore and in 2011, it had a stake of Rs.1554 crore, an almost 70% rise in two years.

Insurance fraud can occur in two ways: Hard insurance fraud and Soft insurance fraud.

Hard insurance fraud: When someone fabricate an accident or injury in order to obtain compensation in an unethical manner.

Soft insurance fraud: When people withhold information from their insurance companies or lie to earn financial benefits from them, this is known as insurance fraud.

At the moment, insurance companies in India are attempting to save costs, and one of the main ways they do so is through detecting fraudulent claims. Diverse tools, such as various data analytics methodologies and different data-mining techniques, are used to detect scammers. To differentiate between legitimate and fraudulent claims, a thorough examination of the fraud claims is necessary. As a result, we use various Machine Learning classification algorithms to detect fraud claims in this study.

The most essential benefit of using a Machine Learning algorithm in the insurance industry is that it makes data sets easier to manage. Machine learning can be applied to structured, semistructured, and unstructured datasets with efficiency. By using superior predictive accuracy, machine learning may be applied across the value chain to identify risk, claims, and customer behaviours. Machine learning has a wide range of potential applications in insurance, ranging from sensitive risk appetite and premium leakage to expense management, subrogation, proceedings, and fraud detection. Machine learning is not a novel method; in reality, it has been around for several generations.

Learning is divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. Since the previous few decades, the majority of insurers have used Supervised Learning to assess risk using known parameters in various combinations to obtain the desired output. Insurers of today are encouraged to engage in unsupervised learning, which has well defined objectives. If any changes are made to the variables, the method detects them and attempts to update them in accordance with the aims. For instance, depending on traffic, the GPS will dynamically suggest different routes based on traffic circumstances. Learning is also used in the insurance business for usage-based insurance.

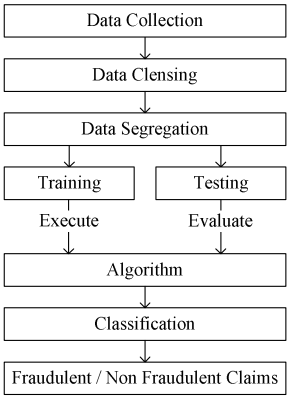


Figure 1

In the typical manner, the support and maintenance of fraud car insurance claim detection is delegated to insurance claim agents or branch inchargers. The claim agent sorts the claims using the facts from the automobile insurance claim and gathers information while waiting for the investigation report. The claim agent determines whether the claim is fraudulent or real based on the information and report gathered.

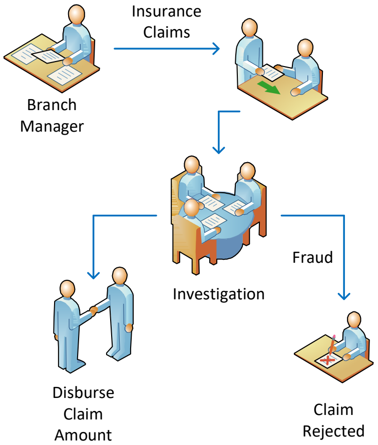


Figure 2 Existing Approach

Figure 1 depicts the typical way of detecting fraud in automobile insurance claims. The claims are assessed by the branch manager using this approach, which is based on the facts collected around indicators. The score is determined by comparing the details of damaged vehicle parts to a checklist. If the score is high, an investigator is assigned to the case to check the damaged vehicle before submitting an investigation report to the in-charger. If the investigation is positive, the claim is considered legitimate and the claim amount is payable. If the answer is negative, the claim is considered fraudulent and will not be paid.

II. LITERATURE SURVEY

Lamberti, et.al [1] proposed an analysis in which smart contract sensor data was used to better understand an on-demand insurance plan. Customers' automobiles were fitted with a mobile application and an electronic tool to generate a prototype. The connection with smart contract was carried out to schedule the automatic updates and collect pictures of automobiles in order to physically modify the policy coverage. The ambient conditions were observed, and the changes were implemented using electronic equipment. The proposed resolution had the potential to reduce policy adjustment costs.

Advantage

- Decreasing the costs of policy change

Disadvantage

- Damages were not identified, nor were smart contracts used to automatically initiate payments.

Kareem, et.al [2] proposed a novel technique for detecting fraudulent claims in health insurance. For the purpose of detecting deceptions, several specific attributes on the claim credentials were examined in order to recognise the association or connection between them. The requirements of the health insurance industry were not completed. As a result, the detection of health insurance deceptions drew more investigators' attention to data mining. As a result, the thriving fortitude of connected attributes might deal with the inconsistency of information in fraudulent allegations, perhaps reducing the chances of fraud in insurance.

Advantage

- Several testing on a real-world dataset clearly verified the predicted scheme's competence.

Kenyon, et.al [3] proposed exploring into big data and data science applications that could be used to predict fraudulent insurance claims. A use-case in an interim insurance firm was used to apply big data, data science, and predictive analytics. For identifying insurance claims fraud, a privacy-protection strategy was proposed.

In addition to the interest of cross-broker and cross-insurer exploitations, the proposed approach used large data samples for policy creation. The results of the tests showed that the rules developed by this methodology were more valuable than those generated by other methods.

Advantage

- As a result, the thriving fortitude of connected attributes might deal with the inconsistency of information in fraudulent allegations, potentially reducing the chances of fraud in health insurance.

Anbarasi, et.al [4] proposed a study related to the fraudulent acquisition of healthcare insurance data. Proactive and retrospective analysis were combined in the proposed strategy. The disconnected nature of anomalous behaviours was controlled by combining reactive and proactive approaches in order to provide a more effective strategy. The suspicious behaviour of a health-related trace was detected using outlier-based predictors for health insurance fraud detection. In the future, further upgraded online fraudulent discovery technologies will be able to identify the nature of health care fraud.

Advantage

- The results of the tests showed that the rules developed by this methodology were more effective than those generated by other methods.

III. SYSTEM MODEL

The proposed research model is a comparative analysis in which several feature selection algorithms are performed and then classification algorithms are built over the selected features to produce the best model that can reliably anticipate fraud claims. The system proceeds in the order depicted in the diagram.

- Data Collection
- Data Pre-processing
- Feature Selection
- Classification
- Result

1. Data Collection

Data collection is the first and most crucial phase in any research project, as all subsequent steps are dependent on the dataset provided. Because the focus of this study is on detecting auto insurance fraud, datasets of auto insurance claims were gathered from online repositories. There are 28,994 samples in the dataset, 32 variables, and one target variable.

2. Data Pre-processing

Data pre-processing entails eliminating missing values, noisy data, and so on. However, there are no missing values or noisy data in the data that was collected. The data has non-numeric categorical values. Because some machine learning algorithms can only work with quantitative data, the data must first be transformed to numeric form.

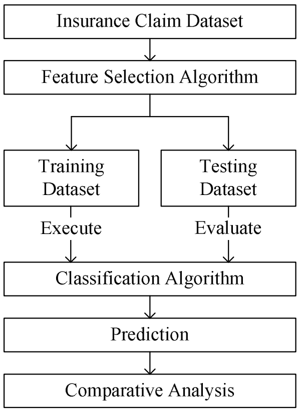


Figure 3 Flow Diagram

3. Feature Selection

After all of the data has been pre-processed, the dataset is subjected to feature selection techniques in order to minimise the data's dimensionality. As is well known, the raw data obtained contains additional information that is rarely employed while developing the model. As a result, data dimensionality reduction is required.

In order to reduce dimensionality, both feature selection and feature extraction strategies are employed, but in this research we only apply feature selection algorithms over the dataset. Subsets of meaningful characteristics are derived from the entire dataset in feature selection.

3.1. L1 Regularization Feature Selection

The selection of L1 Regularization features is based on charging the model with the L1 norm, which result in sparse solutions. The model is linear, and the classifier is Linear SVC. The purpose is to obtain the key features that can be obtained when SelectFromModel of scikit-learn's is utilised.

3.2. Tree-Based Feature Selection

Tree-based feature selection is another feature selection approach that can be used to compute significant features while removing unnecessary ones. ExtraTreesClassifier, an incredibly randomised classifier, is used for feature selection in this module.

3.3 Univariate Feature Selection

This algorithm's operation is based on univariate statistical measures that are used to select the best features. The SelectKBest class scores the features with a function and then removes all but the k highest scoring features.

4. Classification

The significant features are sent into several algorithms for further processing after the feature selection stage. Based on a survey conducted in the field of fraud detection, different categorization methods previously utilised in the field of fraud detection are chosen for processing in this proposed model.

4.1. K Nearest Neighbor

K-nearest neighbours is an instance-based learning algorithm because it stores all of the available examples or classes and then utilises the stored instances to categorise the new instance based on the similarity measurement when it receives a test instance. The new instance is classed using distance functions like Euclidean distance, which can be used to determine which neighbours are closest to the incoming instances. It is well-known for its ease of understanding, simplicity, and quick calculation time. The value of k is important in this approach since it is used to determine the nearest neighbours of a new instance from the stored instances and anticipate the new instance's class. Now, prediction is dependent on the type of problem; for example, if the problem is a classification problem, neighbours are polled and the class with the most votes is chosen. This algorithm does not generate a previous model that is trained in advance; instead, it saves the entire dataset, thus no training is necessary.

4.2. Naive Bayes

Naive Bayes is a supervised machine learning algorithm that is used to predict future class values over a training dataset in which the target classes are known. It's a really effective probabilistic approach. It can be used for multi-class classification as well as binary classification. This approach assumes that the properties in a class are unrelated to other attributes. Because it naively assumes that each attribute is independent of each other given the target variable, this strategy is called "naive." The Bayes theorem is utilised to determine posterior probability in this approach.

4.3. Random Forest

In supervised machine learning, the Random Forest algorithm is the most often used method of learning. For both classification and regression issues, it is a versatile algorithm. You can learn this way by grouping weak learners to construct a more powerful model. Instead of creating single trees as in CART, this algorithm creates a forest of numerous trees. Use of this approach has the advantage of handling missing values and outliers better than other algorithms. It works well with huge datasets with high dimensionality. In order to avoid overfitting the model, each tree contributes its categorization, or we might say that the tree votes for that class.

An instance is categorised as a class with the highest number of votes in a classification problem by using majority voting. In a regression problem, the average of all trees' outputs is taken into account when solving. In terms of statistical modellers, it's like a black box, since they have no way of knowing what the model is going to do.

5. Result

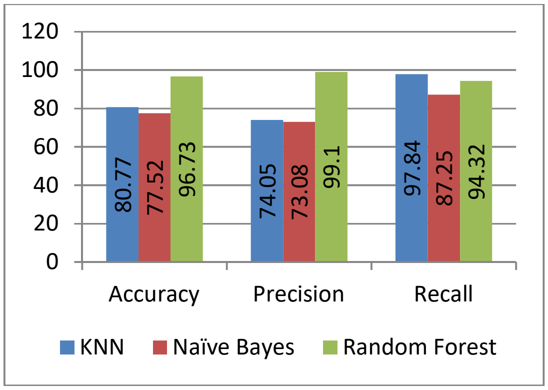
The comparison model is evaluated using performance criteria such as accuracy, precision, and recall. Accuracy is commonly used to measure a model's performance, however it is insufficient, thus we also utilise other performance measures such as Precision and Recall in addition to Accuracy. Precision is a measure of a classifier's exactness, whereas recall is a measure of a classifier's completeness; it primarily deals with the number of correctly classified samples from the total positive samples, hence recall must be taken into account when calculating the classifier's performance.

IV. RESULT

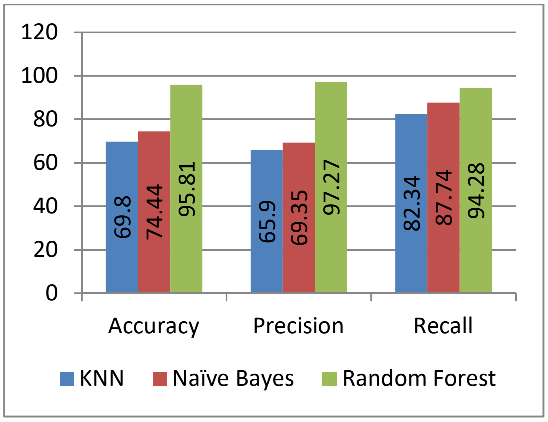
Three feature selection techniques and three classification algorithms have been implemented, as well as performance measurements. After completing the investigation, we can state that Random Forest method delivers 96.73 percent accuracy with L1-Based feature selection algorithm, which is better compared to other classification algorithms. Since we are dealing with fraud detection mechanisms, precision and recall are very important.

After analysing the performance parameters of the comparison model comprising multiple classification algorithms, we can state that different algorithms have a distinct impact on the outcomes. Regardless of which feature selection algorithm is used, Random Forest excels in terms of accuracy and precision. If you're looking for a high recall factor, then the Decision Tree is the finest of all the feature selection algorithms. According to the system requirements, performance parameters should be analysed. Due to the fact that this is a fraud prediction system, false negative situations are unacceptable, which ultimately leads to a high recall factor.

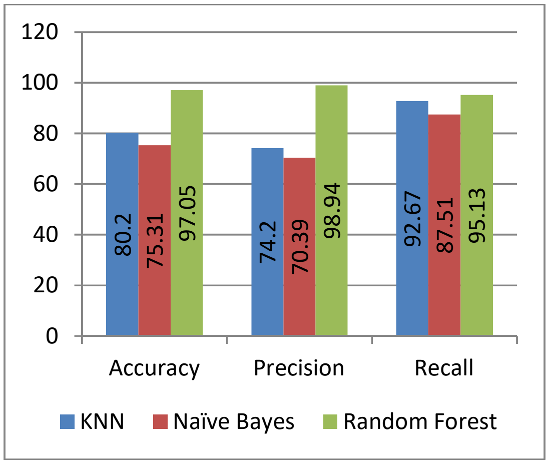
i. Comparison of L1Regularization Feature Selection



ii. Comparison of Tree Based Feature Selection



iii. Comparison of Univariate Feature Selection



VI. CONCLUSION

Comparing Random Forest method to other classification algorithms, based on experimental analysis and overall performance of the model, it can be determined that Random Forest algorithm has given the best and most consistent results. According to the aforementioned experimental results, L1 Regularization based outperforms all other feature selection algorithms, whereas Random Forest and Decision Tree are the top classification algorithms in terms of performance parameters. In terms of accuracy and precision, Random forest algorithm outperforms other classification algorithms, whereas Decision tree method outperforms other classification algorithms in terms of recall.

REFERENCES

[1] Yaqi Li, Chun Yan, Wei Liu, Maozhen Li, “A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification”.

[2] G G Sundarkumar, Ravi V, “A novel hybrid under sampling method for mining unbalanced data sets in banking and insurance”

[3] Maozhen Li, Yaqi Li, Chun Yan, Wei Liu,. “Research and Application of Random Forest Model in Mining Automobile Insurance Fraud”.

[4] Rekha Bhowmik, “Detecting Auto Insurance Fraud by Data Mining Techniques”

[5] H.Lookman Sithic, T.Balasubramanian, " Survey of Insurance Fraud Detection Using Data Mining Techniques"

[6] Tessy Badriyah, Lailul Rahmaniah, Iwan Syarif, " Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance"

[7] Najmeddine Dhieb, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud, " Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations"

[8] Richard A. Bauder, Taghi M. Khoshgoftaar, Aaron Richter, Matthew Herland, " Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims"

[9] Saba kareem, Dr. Rohiza Binti Ahmad, Dr. Aliza Binit Sarlan, " Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining "

[10] Wu Jihong, Wang Junmei, Liu Yanjun, " Design and Research of Insurance Survey Claims System Based on Big Data Analysis"