IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Comparative Analysis on Classification Algorithms of Auto-Insurance Fraud Detection based on Feature Selection Algorithms

Sapna Panigrahi[1*], Bhakti Palkar[2]

[1,2]Department of Computer Engineering, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai-77, Maharashtra India

[*]*Corresponding Author: sapna.rp@somaiya.edu*

Available online at: www.ijcseonline.org

Accepted: 17/Aug/2018, Published: 30/Sept./2018

Abstract— This paper is a comparative analysis of different machine learning algorithms used to detect fraud claims of Automobile/vehicle Insurance claims dataset. In this paper large dataset of automobile insurance claims is used and three feature selection algorithms are applied to the dataset which will be used by the classification algorithms to detect the 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), K-Nearest Neighbor(KNN) and Decision Tree(DT). These algorithms are compared on the basis of performance measures such as accuracy, precision, recall. The proposed model shows that Random Forest works well with respect to accuracy and precision and Decision Tree is the best with respect to recall.

Keywords—Automobile Insurance,Machinelearning, Feature Selection Algorithms, Classification Algorithms.

I. INTRODUCTION

An agreement in which the person pays premiums to an insurance company and in return receives protection from losses is known as an Insurance. In the world there are around 1000 insurance companies and they collect nearly one trillions of dollars each year. If a person makes a false insurance claim so as to get benefit out of it is called as an insurance fraud. The estimated total cost is more than 40 billion in dollars. So it is a big challenge to predict the frauds claims in insurance.

According to a survey of Indiaforensic Research, it is said that the Insurance Sector in India loses around 30,401 Crore of rupees every year because of frauds; in other words it loses 8.5% of its revenues to the frauds. Basically there two parts of Insurance Sector namely Life Insurance and General Insurance. Automobile/Vehicle Insurance is the largest and most profitable sector in General Insurance. Auto Insurance has a stake of Rs.915 Crore in 2009 and Rs.1554 Crore in the year 2011 which is almost 70% of increase in two years.
Insurance frauds occur in two ways:
• Hard fraud: When people report fake accident/injury and obtain the compensation in an illegal way. Such frauds comes under hard fraud.
• Soft fraud: When people hide certain information from their insurance companies or lie to gain financial benefits from company. Such frauds comes under soft fraud.

At present insurance companies in India are working to reduce costs and their main focus in order to reduce costs is by detecting fraud claims. For detection of fraudsters various techniques are used such as various data analytics methods are used, different data-mining techniques Thorough analysis of the fraud claims is must so as to differentiate between genuine claims and fraud claims. So in this paper we are using different Machine Learning classification methods to detect fraud claims.

The rest of the paper is organized as follows, Section II contains the related work done in the field of fraud detection Section III contains the proposed model, flow diagram, different algorithms used in the research, Section IV contains the experimental results obtained in the research work section V concludes research work with future scope.

II. RELATED WORK

Yaqi Li et.al[2017] [1] proposed a prediction model using PCA based Random forest for automobile insurance frauds. But the limitation of this paper is that the algorithm is applied to very few records of the dataset. Hence there is a scope of testing this model over the large set of data.
G G Sundarkumar et al.[2014] [2] in this paper a model is developed which is used to reduce the data imbalance problem in the dataset. The model is created using K-reverse

nearest neighbor along with one class support vector machine (OCSVM).

Maozhen Li et.al[2016] [3] in this paper Random Forest algorithm is used in mining automobile insurance fraud, but the drawback of this paper is that the dataset is small and explanatory variables are less. So the system can be upgraded by using large dataset.
Bhowmik [2011] [4] in this paper authors have used Naïve Bayesian classification network and Decision Tree-Based algorithms classify the auto fraud claims as fraudulent or honest. The model performance was measured using performance parameters and have used Rule-based classification for visualization.

H.Lookman Sithic et.al[2013][5] in this paper financial fraud are detected using some data mining techniques focusing mainly on insurance frauds. But the drawback of this paper is that they have used artificial data so the future work is to use real data to detect fraud.

Adrian Gepp etc.[2012][6] this paper is a comparative analysis of algorithms such as KNN, Decision Tree, Logit Model to predict fraud over automobile insurance fraud dataset.

Clifton Phua et.al.[2014][7] proposed an innovative method which deals with skewed data distributions using the minority over-sampling. In this paper the algorithms used are Back- propagation along with Naive Bayes and C4.5.
Tina R. Patil et.al[8] in this paper Naive Bayes and J48 algorithms are implemented over bank dataset so that sensitivity or true positive rate can be maximized and false positive rate minimized and then compare both the algorithms over some performance parameters.

Xidi Wang et.al.[9] this paper shows a comparison of different classification algorithms such as Decision tree, Neural Network, Bayesian Network, Naïve Bayes and Artificial Immune System. It is concluded that Bayesian Network is better than Neural Network.

Lijia Guo[10] includes informative discovery techniques and data mining techniques on auto insurance claims using cluster discovery methods and decision tree analysis.

III. PROPOSED MODEL

In this paper, the proposed model is a comparative analysis in which different feature selection algorithms are implemented and then classification algorithms are build over the selected features to get a best model which can predict fraud claims accurately. The system follows the steps as shown in the diagram.
A. Data Collection and Pre-processing.
B. Feature Selection Algorithms.
C. Classification algorithms used in the proposed model.

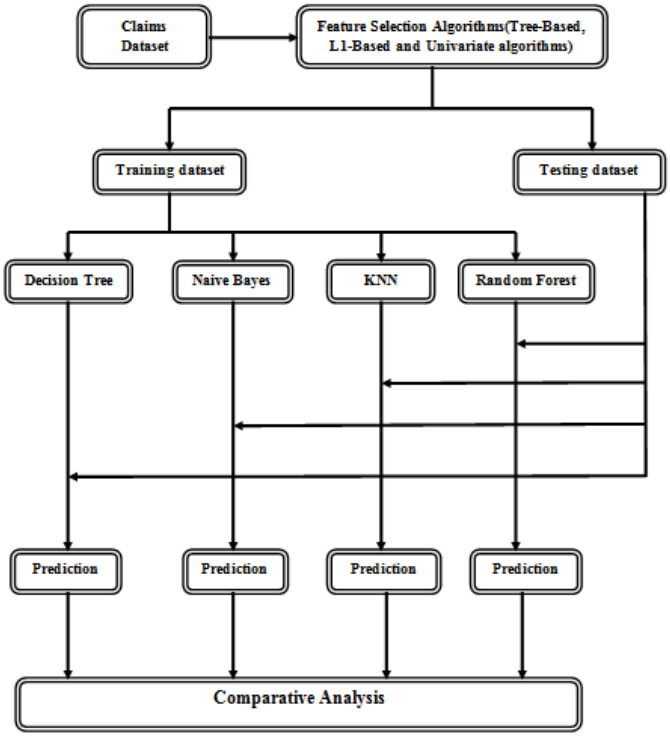

Figure 1: Flow diagram

A. Data Collection and Pre-processing

Data Collection is the first and very important task in any research work to be done as the further steps are dependent on the provided dataset. As in this paper, focus is on Automobile Insurance fraud detection so dataset of Auto Insurance claims are collected from online repositories. The dataset has 28,994 samples and 32 variables and 1 target variable.

Data pre-processing consists of removing missing values, removing noisy data etc. But the data collected do not contain any missing values or noisy data. The data has categorical values which are non-numeric. Some machine learning algorithms are feasible with numeric data so the data is converted into numeric form.

Once all the samples are in numeric format we split the data into Training dataset and Testing dataset in the ratio 70:30 which means that 70% of data is used to train the model and 30% is used to test the correctness of the model in predicting the claims.

B. Feature Selection Algorithms

After all the data is pre-processed, further feature selection algorithms are applied to the dataset so as to reduce the dimensionality of the data. As it is known that the raw data

collected consists of extra information which are mostly not used while building the model. So there is a need for dimensionality reduction of the data. Both feature selection and feature extraction algorithms are used in reducing dimensionality, but in this paper we are using only the feature selection algorithms over the dataset.

In feature selection subsets of meaningful features are derived from the whole dataset. In this paper, we are using scikit-learn library for feature selection. These are the following feature selection algorithms used to build the model.

a) Univariate Feature Selection.

The working of this algorithm is based on univariate statistical measures which is used to select the best features. The SelectKBest class scores the features using a function and then removes all but the k highest scoring features.

b) L1-based Feature selection.

L1- based feature selection is based on penalizing the model with L1 norm which gives sparse solutions. The model is linear which uses LinearSVC as a classifier. The goal is to get the important features which can be achieved when used along with SelectFromModel of scikit-learn.

c) Tree-Based Feature Selection.

Tree-based feature selection is also a feature selection algorithm that can be used to compute important features, which ultimately eliminates the irrelevant features. In this module ExtraTreesClassifier is used for feature selection which is an extremely randomized classifier.

Each Feature Selection are used over the dataset individually to get the subset of important features and then Classification algorithms are trained over the selected features to get the results.

C. Classification Algorithms used in this model

After the feature selection step is done the important features are fed for further processing to different algorithms. Based on the survey done in fraud detection, different classification algorithms already used in the field of fraud detection are chosen for processing in this proposed model.

a) ***Decision Tree Algorithm***:

Decision tree is one of the most popular supervised machine learning algorithms which can be used for both classification and regression problems. It works for both categorical and continuous input and output variables. The dataset is divided in smaller sets and a tree is created simultaneously. The resultant tree created consists of nodes. The nodes in the tree are root node, decision nodes and leaf nodes. The tree starts from the root node, if a node is divided into two or more branches it is a decision node and if a node cannot be further sub-divided it is a leaf node. Leaf node represents class labels and attributes in the internal node of the tree. Decision tree builds the tree by splitting the samples into two or more homogeneous sets. The splitting is done based on the information gain and entropy. Depending on the information gain the attribute is selected to be further explored. To get the information gain it is necessary to get the entropy of each variable which can be calculated with the expression given below:

$$\text{Entropy}(S) = - p \log_2 p - q \log_2 q$$

$$\text{Gain}(S, V) = \text{Entropy}(S) - \sum (|S_v|/|S|) \text{Entropy}(S_v)$$

Where, p is the proportion of positive class samples in S

q is the proportion of positive class samples in S

S is a collection of positive and negative instances

V is an attribute whose information gain is to be calculated.

b) ***Naive Bayes Algorithm:***

Naive Bayes (NB) is a supervised machine learning method that is implemented over training dataset in which the target classes are known which is used to predict the future class value. It is a powerful probabilistic method. It is used for binary classification as well as multi-class classification. The assumption in this algorithm is that the attributes in a class is not related to other attributes. This technique is named "naive" because it naively assumes that each attribute is independent of each other given the target variable. This algorithm is based on Bayes theorem which is used to calculate posterior probability P(h/D). Mathematical expression of Bayes Theorem is given below.

$$P(h/D) = [P(D/h) P(h)] / P(D)$$

Where, P (*h/D*) is Probability of *h* given *D*, also called Posterior Probability.

P (*D/h*) is Probability of *D* given *h*,

P(*h*) is probability of hypothesis,

P(*d*)is probability of training data

c) K-Nearest Neighbor:

K-nearest neighbors is an instance based learning algorithm since it stores all the available instances or classes and then when it gets a test instance it uses the stored instances to classify the new instance which depends on the measurement of the similarity. The new instance is classified based on some distance functions such as Euclidean distance which can be used to find the neighbors which are closer to the incoming instances. It is well-known for its simplicity, ease of interpretation and low calculation time. The value of k plays a vital role in this algorithm as k value is used to find the nearest neighbors of new instance from the stored instances and predict the class of the new instance. Now, prediction depends on the type of problem, if it is classification problem then voting is done among its neighbors and the majority voted class is chosen. In this algorithm there is no prior model created which is trained in

advance rather it stores the whole dataset, hence there is no learning required.

d) Random Forest Algorithm:

Random Forest algorithm is the most popular supervised machine learning algorithm. It is the versatile algorithm which is used for both classification as well as regression problems. It is a type of ensemble learning method in which the weak learners come together to create a powerful model. In this algorithm forest of multiple trees are created instead of single trees as in CART. The main advantage of using this algorithm is that it can handle missing values, outlier values compared other algorithms. It can work with large datasets with higher dimensionality very well. As it grows multiple trees in the forest, so it won't overfit the model each tree gives its classification or we can say that the tree votes for that class. In classification problem, to classify a new instance majority voting is considered and the new instance is classified as the class with most number of votes. In regression problem, the average of output from all trees are considered. This algorithm acts like a black box for statistical modelers since there is no control over what exactly the model does.

IV. EXPERIMENTAL RESULTS

The analysis of the proposed model is done on the basis of some performance parameters such as Accuracy, Precision, Recall. Generally Accuracy is used to measure the performance of a model but it is not enough, so we use more performance parameters such as Precision and Recall along with Accuracy. These parameters help us to check the robustness of our model whether it can predict the unknown data samples correctly. All the performance measures depends on Confusion Matrix which is build based on the prediction of dataset made by the model. Precision is nothing but the measure of classifier exactness while Recall is a measure of a classifiers completeness, it mainly deals with number of correctly classified samples from the total positive samples and hence recall must also be consider to calculate the performance of the classifier.

The Experimental results are shown in the Table1 below, all values are in percentage it has 3 feature selection algorithms and 4 classifications algorithms implemented and the results of performance measures. After all the analysis we can say that with respect to accuracy measure over all three feature selection algorithms, Random Forest algorithm gives 96.73% with L1-Based feature selection algorithm which is better compared to other classification algorithms. But this is not enough since we are dealing with fraud detection mechanism we are considering precision and recall. With respect to precision measure, Random Forest algorithm gives 99.10% with L1-Based feature selection algorithm which is better compared to other classification algorithms. Now with respect to recall measure, Decision Tree works well with L1-Based Feature Selection which gives 95.50%.

Table 1: Comparison results

Algorithms	Decision Tree Algorithm			Naive Bayes Algorithm			K-Nearest Neighbor			Random Forest Algorithm		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
Tree-Based Feature Selection	94.48	94.32	94.67	74.44	69.35	87.74	69.84	65.92	82.34	95.81	97.27	94.28
L1-based feature selection	94.60	93.84	95.50	77.52	73.08	87.25	80.77	74.05	94.84	96.73	99.10	94.32
Univariate Feature Selection	95.21	97.57	95.95	75.31	70.39	87.51	80.20	74.20	92.67	97.05	98.94	95.13

Tree-Based Feature Selection algorithm: The algorithm selects total 17 features out of 32 features of the whole dataset and then the classification algorithms are trained and tested over the selected features. The results using the algorithm are shown in Figure1 below, considering the performance parameters with respect to accuracy and precision both, Random Forest is the best with 95.81% and 97.27% respectively when compared with other classification algorithms. And with respect to recall parameter Decision Tree gives 94.67% and Random Forest gives 94.28% which shows a slight difference.

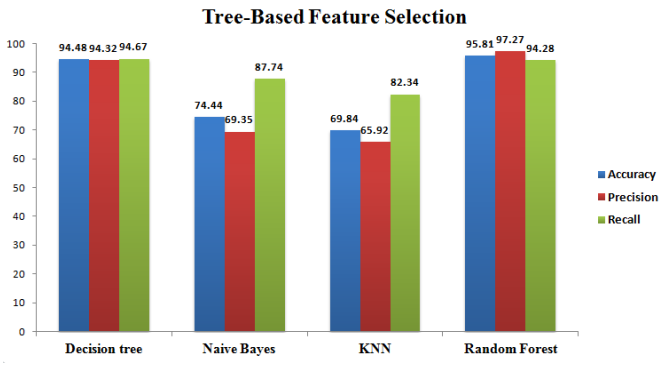

Figure 2: Tree-Based feature selection

L1-Based Feature Selection Algorithm: This algorithm selects 28 features out of 32 features of the dataset and then the classification algorithms are trained and tested over the selected features. The results using the algorithm are shown in Figure2 below, considering the performance parameters with respect to accuracy and precision both Random Forest is the best with 96.73% and 99.10% respectively when compared with other classification algorithms. And with respect to recall parameter Decision Tree gives the best result which is 95.50%, KNN and Random Forest gives 94.84%and

94.32% respectively which shows a slight difference between the results

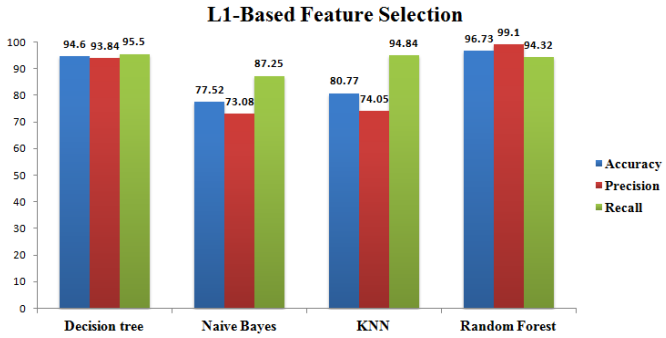

Figure 3: L1-Based feature selection

Univariate Feature Selection: This algorithm selects 15 features out of 32 features of the dataset and then the classification algorithms are trained and tested over the selected features. The results using the algorithm are shown in Figure3 below, considering the performance parameters with respect to accuracy and precision both Random Forest is the best with 97.05% and 98.94% respectively when compared with other classification algorithms. And with respect to recall parameter Decision Tree gives 95.95% which is best among all algorithms.

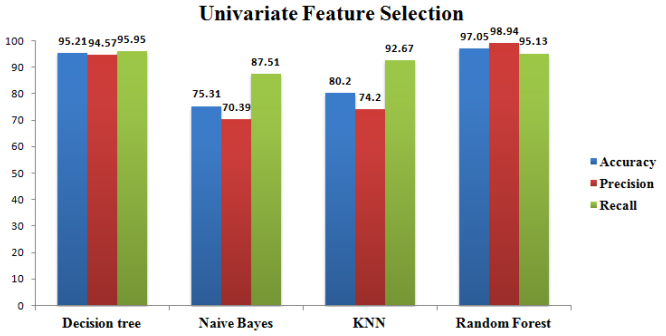

Figure 4: Univariate feature selection

After the analysis of the proposed model including different classification algorithms in terms of the performance parameters we can say that different algorithms have a different impact on the results. In terms accuracy and precision, Random Forest excels irrespective of the feature selection algorithm. In terms of recall, Decision Tree is the best for all the feature selection algorithms this means that it excels in giving high recall factor. Now, depending on the system requirements the performance parameters should be considered for analysis purpose. As this is a fraud prediction system, the false negative cases are not acceptable which ultimately leads to recall factor so system should have high recall also along with accuracy and precision.

V. CONCLUSION and Future Scope

From the experimental analysis, considering the overall performance of the model, it can be concluded that Random Forest algorithm have shown the best and most consistent results with L1-Based feature selection algorithm compared to other classification algorithms. As per the above experimental results, it shows that among all feature selection algorithms used in this paper, L1-Based exceeds the other feature selection algorithms and among all classification algorithms, Random Forest algorithm and Decision Tree algorithm are the best algorithms which gives the best results with respect to performance parameters. Now considering the performance parameters for comparison, based on accuracy as well as precision Random forest algorithm exceeds other classification algorithms giving the high result and based on recall parameter Decision tree algorithm gives high results.

The future scope of this research work can be further considering the time complexities of algorithms as it is not considered at present in this paper. Also performance of some unsupervised machine learning algorithms for fraud detection can also be examined over the same dataset used in this paper.

ACKNOWLEDGMENT

The author is thankful to Prof. Bhakti Palkar of K. J. Somaiya College of Engineering for her guidance and suggestions.

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". Applied Soft Computing Journalhttp://dx.doi.org/10.1016/j.asoc.2017.07.027

[2] G G Sundarkumar, Ravi V, "A novel hybrid under sampling method for mining unbalanced data sets in banking and insurance" [J]. Engineering Applications of Artificial Intelligence, 2015: 368-377.

[3] Maozhen Li, Yaqi Li, Chun Yan, Wei Liu,. "Research and Application of Random Forest Model in Mining Automobile Insurance Fraud". 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) 2016.

[4] Rekha Bhowmik, "Detecting Auto Insurance Fraud by Data Mining Techniques", Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, APRIL 2011

[5] H.Lookman Sithic, T.Balasubramanian, " Survey of Insurance Fraud Detection Using Data Mining Techniques" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-3, February 2013

[6] Adrian Gepp, J. Holton Wilson, Kuldeep Kumar and Sukanto Bhattacharya, "A Comparative analysis of Decision Trees and other computational datamining techniques in automotive insurance fraud detection", Journal of Data Science 10(2012)

[7] Clifton Phua, Damminda Alahakoon, and Vincent Lee, " Minority Report in Fraud Detection: Classification of Skewed Data", ACM SIGKDD Explorations 2014.
[8] Tina R. Patil, Mrs. S. S. Sherekar, " Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification ", International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013.
[9] Manoel Fernando Alonso Gadi, Xidi Wang and Alair Pereira do Lago, "Credit Card Fraud Detection with Artificial Immune System "
[10] LijiaGuo,"Applying Data Mining Techniques in Property~Casualty Insurance "
[11] M. Shukla, A. K. Malviya, "Analysis and Comparison of Classification Algorithms for Student Placement Prediction", International Journal of Computer Sciences and Engineering Vol.-6, Issue-6, June 2018.

Authors Profile

Miss Sapna Panigrahi pursued Bachelor of Engineering in computer engineering from A.C Patil College of Engineering, Mumbai, Maharashtra, India in 2016. She is currently pursuing M.Tech in computer engineering from K.J Somaiya College of Engineering, Mumbai, Maharashtra, India. Her research interest is in Machine Learning, Data mining.

Mrs. Bhakti Palkar is currently working as an Associate Professor in K.J Somaiya College of Engineering, Mumbai, Maharashtra, India.

