Algorithms

Kamran,
PG Scholar/Data Science,
Alliance College of Engineering and Design,
Alliance University,
Bangalore, India.

Soumyakshya Das,
PG Scholar/Data Science,
Alliance College of Engineering and Design,
Alliance University,
Bangalore, India.

Dr. Senbagavalli M
Associate Professor/IT,
Alliance College of Engineering and Design,
Alliance University,
Bangalore, India.

Abstract - Insurance Fraud is a grave, earnest and a growing problem and there is a universal recognition that the traditional approaches to tackling the problem of fraud is deficient. An alternative methodology is to comprehend, understand and optimize existing practices in the detection of fraud. This paper presents an ethnographical study through the method of machine learning classificational algorithms such as KNN, SVM, Random Forest, XG Boosting and Decision Tree to explore the nature of fraud detection in a leading insurance company. The result of the study suggests that an occupational focus on the practices of fraud detection can complement and enhance forensic and data-mining approaches to the detection of potentially claims. This paper presents an alternative, though still embryonic, a way to get a verdict whether a claim is deceitful or not using k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, Decision Trees and XG Boosting such classification techniques.

Keywords - *Insurance Fraud, Decision Trees, Support Vector Machines, k-Nearest Neighbors and XG Boosting.*

I. INTRODUCTION

An approximated 10 percent (%) of claims filed with the Global insurance industry are duplicitous and fraudulent. Typically, only a single digit percentage of the total claims are prohibited or recuperated as part of claim handling fraud investigation units [1]. Insurance artifice is increasingly exponentially with the easy access of modern technology and communication, resulting in the loss of trillions of dollars worldwide each year. Insurance fraud is a worldwide problem. Handling fraud manually has always been expensive for insurance companies. Data analytics provides an effective way to be more proactive in the fight against fraud and to identify transactions that indicate fraudulent activity or the heightened risk of fraud.[2] According to Bolton & Hand, the appropriate overall strategy for fraud discernment is to use a compartmentalized system of investigation.[3] Accounts with very high suspicion scores merit immediate and intensive (and expensive) investigation, while those with large but lower scores merit closer (but not expensive) observation. Techniques for fraud detection are important if we are to identify fraudsters once fraud prevention has failed. In this paper, we will apply statistical hypothesis testing techniques. With an increase in financial accounting fraud in the current economic scenario experienced, financial accounting fraud detection has become an emerging topic of great importance for academical research, and industries.[4]

II. MOTIVATION

The challenge we are tackling in our work is to provide a faster and more accurate result which in case of fraudulent insurance claims. It is now possible for test samples being collected from home through the medium of various sites but over here we have received the dataset from insurance companies. 3 Key Drivers towards EDA (Exploratory Data Analysis):

- Gaining valuable hints for Data Cleaning
- Ideas for Feature Engineering
- Get a "feel" for the dataset, which will help quantifying inferences and deliver greater impact.

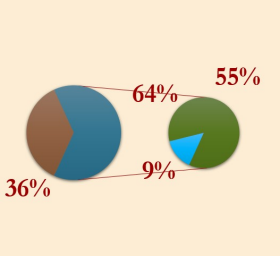

Fig.1 Exploratory Data Analysis

The dataset has a total of 12000 observations with 77 variables:

- 49 has numerical variables (64%)
  - -7 are discrete (9%)
  - -42 are continuous (55%)
- 28 are categorical data (36%)
- Target Variable 'Fraud' was of categorial type

This paper and the algorithm provide an alternative and a confirmation of assurance that the results which shows the insurance target claims are correct or not. Also, to show which algorithms work the best and give the most optimized results.

III.LITERATURE SURVEY

[2] Yaqi Li et.al[2017] proposed a prediction model using PCA based Random forest for automobile insurance frauds and limitation of this paper is that the specific 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] [3] 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] used Random Forest algorithm 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] [5] 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][6] detected financial fraud 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][7] 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] combined 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] shows a comparison of different classification algorithms such as Decision tree, Neural Network, Bayesian Network etc., It is concluded that Bayesian Network is better than Neural Network. Senbagavalli M[2020][2018][10][11]

proposed a Reliability System for Filtering Malicious Information on Social Network and Implementation of Online Crime Reporting System.

A. *KNN*

k-nearest neighbor is a non-parametric algorithm which is used for classification and regression. The basic functionality of knn is based on the weight/distance calculated and then put in the algorithm to differentiate between data points. It does so by grouping those close to the set weight as one class or type of data.

B. *SVM*

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. When provided with a labelled training data, SVM outputs an optimal hyperplane which categorizes new examples.

C. *XG Boosting*

XG Boost is a decision tree-based ensemble algorithm which uses a gradient boosting framework. XG Boost is applicable in a variety of problem statement from classification, ranking and user-defined prediction.

D. *Random Forest*

Random forest is an ensemble learning method used for classification and regression, it operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

E. *Decision Tree*

A decision tree is a predictive model expressed as a recursive partition of the feature space to subspaces that constitute a basis for prediction. Nodes in decision trees with outgoing edges are the internal nodes, while the other nodes are terminal nodes or leaves. Decision tree classifies using a set of hierarchical decisions on the features. The decisions made at internal nodes are the split criterion. In decision trees, each leaf is assigned to one class or its probability. Small variations in the training set results in different splits leading to a different decision trees, which makes it more versatile. Senbagavalli, M &TholkappiaArasu,G[2016][12] proposed Decision Tree based Feature Selection algorithm for Cardiovascular Disease.

IV.COLLECTION OF DATA

The data was predominantly good and missing value of treatment was done by following ways:

- Missing values for different variables were detected:
  - BMI: 0.0142%
  - CLI_INCOME: 0.0051%
  - POL_INIT_PREM_PAY_MTHD: 0.0268%
  - POL_MED_NMED_CHK_IND: 0.0005%
  - EMR: 0.0221%
- Missing variables treatment
  - **Categorical data:** Replaced with new label.
  - **Numerical data:** Replaced missing values with median.

Data Preparation was done by taking following measures:

- Converted categorical data into numeric.
- This was done by grouping the categorical variables with respect to target variable and replacing results with their mean value with help of label encoder.

- **Random Forest Classifier (Embedded Method)** was used for feature selection due to below reasons.

  A) They are highly accurate

  B) They generalize better

  C) They are interpretable

- Models were built on both Up sampled and Under-sampled datasets.
- Datasets were divided into train and test.

V. IMPLEMENTATION

Model Validation- Decision Tree

- Decision tree algorithm falls under the category of supervised learning
- Feature values are preferred to be categorical.
- We have run decision tree on both up sampled and under- sampled datasets.
- In this model we plotted a correlation plot which was used to reduce the features in DT model, lower correlation variables were removed.
- We also worked on **GridSearchCV** to find best hyperparameters.

Hyperparameter Tuning

- The following parameters were tuned to fit the best model:
- max_depth, min_samples_split, min_samples_leaf.

Fig.2 Decision Tree

Feature suggested by RF & correlation plot.

Features with low or negative correlation were removed.

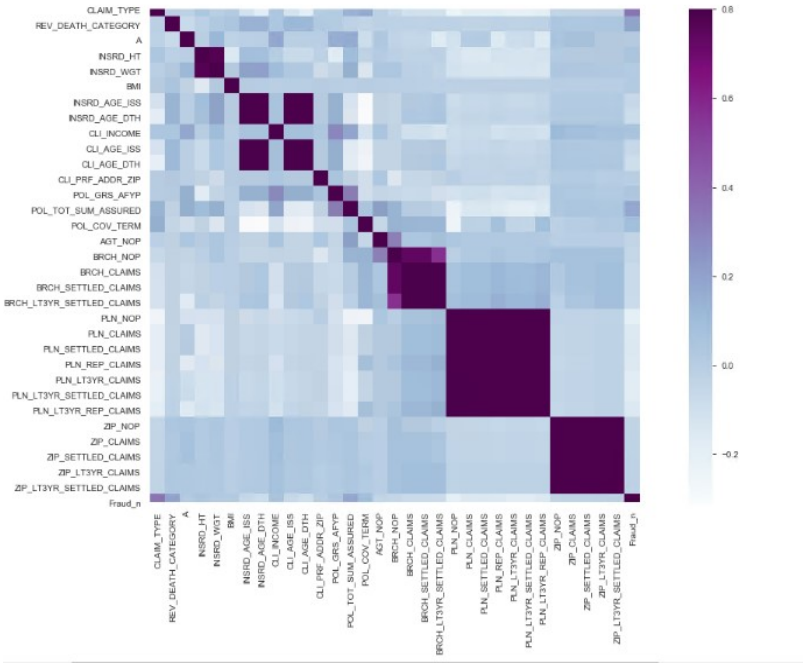

Fig.3 Correlation plots

A. Confusion Matrix

After getting the overall accuracy of about 97%, a plot for confusion matrix was designed to give it a visual aid.

A confusion matrix is a table that describes:

- true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- true negatives (TN): We predicted no, and they don't have the disease.
- false positives (FP): We predicted yes, but they don't actually have the disease.
- false negatives (FN): We predicted no, but they actually do have the disease.

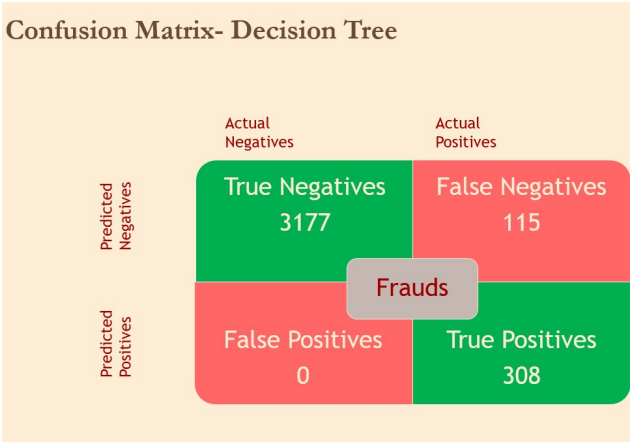

Fig.4 Confusion Matrix for Decision Tree

Model Validation- Random Forest

- The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one "forest."

- It provides higher accuracy.
- It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks)

We also performed **RandomSearchCV for RF** to find best hyperparameters or hyper tuning.

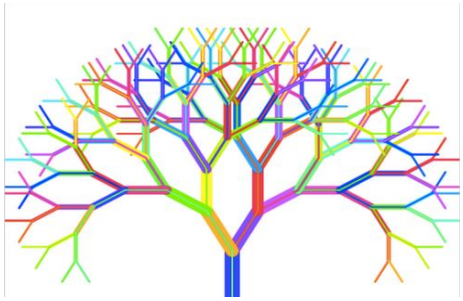

Fig. 5 Random Forest Model Validation

B. Confusion Matrix

After getting the overall accuracy of about 99%, a plot for confusion matrix was designed to give it a visual aid.

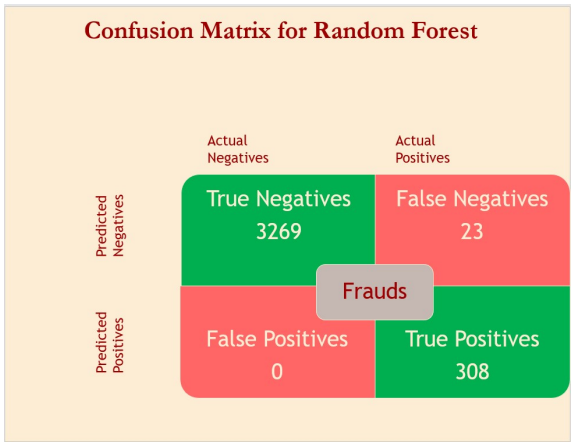

Fig. 6 Confusion Matrix for Random Forest

Model Validation- KNN

- The KNN model predicts the values based on the weighted of surrounding **K nearest** neighbor.
- It can be used for both classification and regression tasks
- As you increase the number of nearest neighbors, the value of k, accuracy might increase.
- In this model we split Train on under sample data balanced data frame & Test split on unbalanced data frame keeping variables based on feature selection output.
- Train Split on Up sampled balanced by keeping variable based on feature selection output.

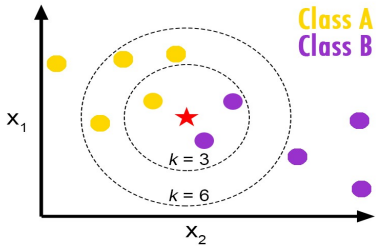

Fig.7 Model Validation-KNN

C. *In KNN, finding the value of k is not easy..*

Simple method to determine k = sqrt(n)

Here showing plotted accuracy vs K-value graph to determine K

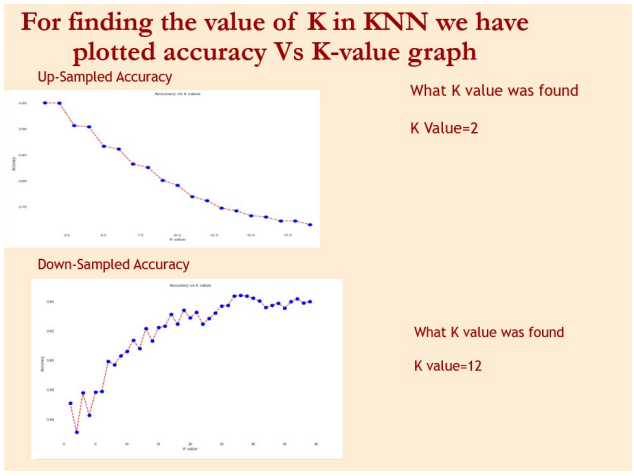

Fig.8 Finding the value of K in KNN

Performance Metrices of KNN:

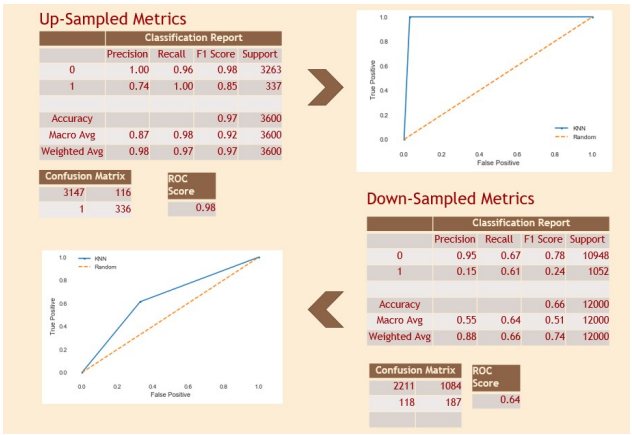

Fig9. Showing performance metrices of KNN

Performance Metrices of XG Boosting

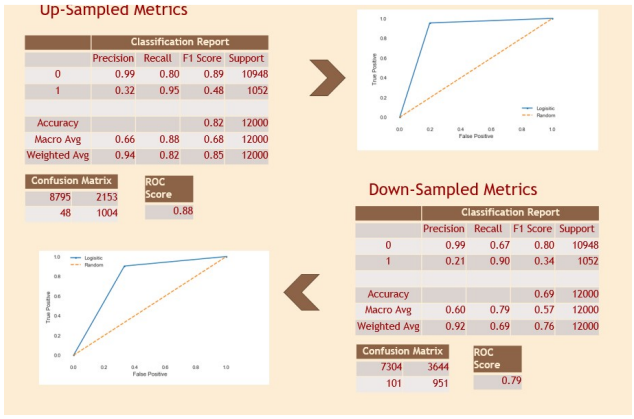

Fig.10 Performance Metrices of XG Boosting

Performance Metrices of Support Vector Machines

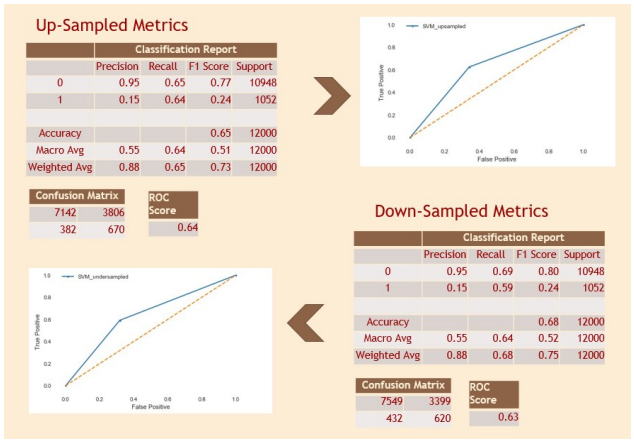

Fig.11 Performance Metrices of Support Vector Machines

Performance Metrics for Logit

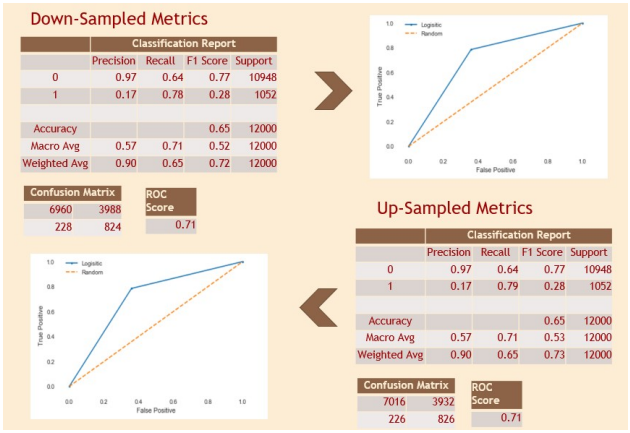

Fig.12 Performance Metrices of Logit

Models	Accuracy	Precision	Recall
Decision Tree Classifier	97%	98%	97%
Random Forest	**99%**	**99%**	**99%**
KNN	97%	98%	97%
XG Booster	82%	94%	82%
SVM	65%	88%	65%
Logit	65%	90%	65%

Fig.12 performance comparison of different classification algorithms

The algorithms ability to identify true results also serves as a solid finding that machine learning and precisely Random Forest can be used in real life scenarios, provided with a real life dataset. The following features found to be impacting the Fraud the most and we recommend every future claim to be passed through our recommended RF model.

- Policy Sum Assured – We recommend to thoroughly examine High Sum assured policies as the chances of Fraud are higher.
- Branch Hist policy sold count- Relatively New Branch with less Historical Policy Counts have more possibility of Fraud.
- Claim Type – Claims Filed within 0-1 Year should be thoroughly scrutinized.
- Client Income – Lower Income Group have more possibility towards Fraud.
- Plan historical sold count – Newer Plans are more Prone towards Fraud.

VI.CONCLUSION

As one can gather from the above documentation of our conducted experimentation, the algorithm designed to work on Insurance Fraud the best accuracy, precision, and recall (99%) is Random Forest.

VII.FUTURE WORK

The implementation of such an algorithm might work better on other forms or other such relevant work. It can be only determined once the experiment is carried on using proper datasets of such kind. The algorithm can be a great in insurance fraud detection also detection in anomalies. Artificial Intelligence is also a big part to play in detection of insurance fraud. MLP would be likely to be used in the future for detection of such cheats or frauds in the future.

References

[1] Centres for Disease Control and Prevention. Evaluating and Testing Persons for Coronavirus Disease 2019 (COVID-19).

[2] 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

[3] 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.

[4] 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 (ICNCFSKD) 2016.

[5] 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

[6] 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

[7] 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)

[8] A. Shen, R. Tong & Y. Deng, (2007) "Application of Classification Models on Credit Card Fraud Detection," Service Systems and Service Management, pp2-5.

[9] W.-H. Chang & J.-S. Chang, (2012) "An effective early fraud detection method for online auctions," Electronic Commerce Research and Applications, pp346-360.

[10] Senbagavalli, M,'A Reliability System for Filtering Malicious Information on Social Network', International Journal of Management, Technology And Engineering,volume 10,issue I,pages 202-214, 2020.

[11] Senbagavalli, M, 'Implementation of Online Crime Reporting System',International Journal of Research Review In Engineering And Management, Volume -2,Issue -4,Apirl -2018,Page No:1-13, 2018.

[12] Senbagavalli, M &TholkappiaArasu,G , 'Opinion Mining for Cardiovascular Disease using Decision Tree based Feature Selection' , Asian Journal of Research in Social Sciences and Humanities –Vol. 6, No. 8, August 2016, pp. 891-897, ISSN 2249-7315,2016.