AutoInsurance Fraud Detection using Machine Learning

<Student Name>

<Student email>

University of North Texas

Professor: <Professor Name>

Introduction

According to the *Federal Bureau of Investigation*, non-health insurance costs an estimated loss of $40 Billion US dollars per year, or about $400 to $700 per U.S family in annually increased premiums. Such numbers of currency represent money that actually goes to fraud from the pockets of innocent people, who seek safety and financial help at time of crisis.

Automobile insurance fraud is a frequently occurring fraudulent activity; according to a *study by Versisk*, auto insurers take a loss of at least $29 Billion US dollars per year! This total is a direct result of fraudulent activities that involve drivers.

All of these add up to the direct increase in premiums for other policy holders. This problem is usually addressed with a combination of various methods and techniques, in order to determine any fraudulent activity that might have taken place, such as:

- *Claims Investigation*: The insurance companies usually have teams of investigators who review and investigate the claim on ground. These teams look for inconsistencies within a filed claim, by interviewing witnesses, examining the vehicles and site at which the incident had occurred, additionally reviewing the police reports and performing extended research on their own.
- *Past Claims and Information Sharing*: Insurance companies often share information about known or suspected fraudsters identified by law enforcement agencies, such as the National Insurance Crime Bureau’s (NCIB) datasets in the United States.
- *Auditing and Review*: Insuring companies also conduct regular audits and additional reviews of claims, and underwriting processes to identify potential frauds and risks.

These methods and techniques are both expensive to conduct and time consuming which requires cooperation between various parties involved. This makes it less than ideal in order to potentially detect an insurance fraud sooner rather than later. Using Machine Learning Algorithms, I intend to train a model that can identify inconsistencies for a filed claim, when this is paired with the above activities the overall time that might be taken to detect and report a fraud will be reduced significantly.

The proposed idea is to have a Machine Learning model: Decision Tree Learning Algorithm, which undergoes supervised learning over the dataset of previously filed claims, and has a target class of *fraud_detected*. This simple solution can be used by insurance companies to work on a statistical estimate and proceed with the necessary activities to verify this hypothesis of fraud, which is a result of my model. Rather treating every case the same way which would be a waste of both time and resources.

Literature Review

2.1 Use of Optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection

This is a research paper, published by Authors: *Sharmila Subudhi* and *Suvasini Panigrahi*, which aims to provide a hybrid approach to detect fraud in automobile insurance claims using a combination of Genetic Algorithm based Fuzzy C-means clustering and supervised classifier models. They first took a portion of the insurance data as a test set, leaving the rest as a training set. They applied the FCM clustering to the training set to create groups of similar data. Then, they used these groups to classify the test set data into three categories: genuine, malicious, and suspicious.

They discarded the genuine and malicious data and focused on the suspicious cases. These cases were analyzed using four different classifiers: Decision Tree, Support Vector Machine, Group Method of Data Handling, and Multi-Layer Perceptron. They used a technique called 10-fold cross-validation to train and validate their models.

By conducting experiments on a real-world car insurance dataset, they demonstrated that their proposed method effectively detects automobile insurance fraud.

2.2 Detecting insurance claims fraud using machine learning techniques

This research paper, is a good reference to the work that I intend to do, this paper authored by *Riya Roy*, and *Thomas George K.* aims to enhance the detection process by utilizing machine learning techniques. Machine learning allows computers to learn from data and make predictions or decisions without being explicitly programmed to do so. This approach can potentially improve the accuracy and efficiency of detecting fraudulent claims.

The authors compare the performance of various machine learning techniques such as Decision Tree and Random Forest by calculating a confusion matrix, which is a table used to describe the performance of a classification model on a dataset for which the true values are known. The confusion matrix helps determine important evaluation metrics such as accuracy, precision, and recall. The findings may contribute to improving the detection process and reducing the financial impact of fraudulent claims on the insurance industry.

However, I intend to determine hypermeters that are best suited for Decision Tree and other learning algorithms, also draw contrast among these to identify the best possible solution in order to classify the label “fraud”.

2.3 A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification

This paper authored by: *Yaqi Li, Chun Yan, Wei Liu,* and *Maozhen Li* presents a new and improved method for detecting car insurance fraud by combining three techniques: Random Forest, Principle Component Analysis (PCA), and Potential Nearest Neighbor. The goal is to make fraud detection more accurate and reduce errors. The method is tested against other similar techniques and shows better results. The authors also claim that it is also applied to real car insurance fraud cases to find fraud patterns.

The authors use PCA to change the data at each step when deciding how to split it, which makes the trees in the Random Forest more diverse and improves accuracy. The relationship between Random Forest and Adaptive Nearest Neighbors is studied, and a new voting method based on Potential Nearest Neighbor is proposed to replace the traditional majority vote.

They tried to test this new method by using 12 different data sets from various fields. And then compared these results to those generated by other methods like Oblique Decision Tree Ensemble, Rotation Forest, and basic Random Forest. The authors, conclude that the proposed new method is less prone to errors, and that it is a newer approach for detecting automobile insurance fraud, without losing sight of important aspects of the case.

2.4 Modeling Insurance Fraud Detection Using Ensemble Combining Classification

This research paper is authored by Amira Kamil and Ajith Abraham focuses on improving insurance fraud detection (IFD) models by utilizing ensemble combining classifiers. The paper builds upon previous work that addressed imbalanced datasets using a novel partitioning-under sampling technique and designed base-classifier IFD models using Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN).

The authors propose innovative IFD models by applying ensemble combining classifiers, specifically Grading, Stacking, and Vote, on the existing base-classifier models (IFDDT, IFDSVM, and IFDANN). The research employs ten-fold cross-validation for testing, and the results indicate that the DTIFD model slightly outperforms the other proposed models.

They additionally perform some experiments: within the experiments, three ensemble combining classifiers were applied: Grading: Grading improved the IFDANN model from 87.7% to 89.5%. However, IFDSVM and IFDDT did not show improvement, with recall values decreasing to 93.4% and 94.3% respectively. The best model using Grading was the ensemble combination of IFDSVM and IFDDT, with a recall of 95.4%. Stacking: Stacking improved the IFDANN model from 87.7% to 94.3%, but did not improve IFDSVM and IFDDT models. The best model using Stacking was the ensemble combination of IFDANN and IFDDT, with a recall of 94.6%. Vote: Vote improved the IFDANN model from 87.7% to 92.6%, but didn't improve the other two models. The best model using Vote was IFDDT, with a recall of 94.3%.

In conclusion, the research proposes innovative IFD models using ensemble combining classifiers Grading, Stacking, and Vote, on existing base-classifier models. The results reported by the authors, suggest that the DTIFD model performs slightly better than other proposed models, with ensemble combining classifiers improving the performance of some models.

Data Exploration

For the classification task that I intend to perform, I identified an open source dataset available on the platform Kaggle. This dataset titled, “Insurance Fraud Detection” was posted by a user named Arpan129 with the latest update made to this data set in late 2022.

This dataset is collected from a project at Indian Institute of Management Calcutta. This dataset is very rich in features, as it contains close to 39 variables and a target variable “fraud reported”.

It is worth mentioning that the dataset only has 1000 data points (rows) in it, however the extensive number of features, balance the lack of lager corpora, out of these 38 features (excluding the target variable.) 18 of these are numerical type (continuous) and 19 features are categorical in nature.

Upon initial over view of this dataset, I realized it also needs some data cleaning and normalization as most of these categorical features contain data of string type. Additionally I could also perform some feature engineering to reduce the variance of each feature, which would bring down the complexity of the tree, resulting in a more interpretable data that can be used for training, for example: the feature that describes the automobile model can be co-related to the cost incurred, as is the case that if the model of automobile is higher than ordinary cars, then the cost incurred to cover the damages could be higher. Which could be a reason why someone might want to commit an insurance fraud i.e. higher return on investment in terms of capital invested in coverage. This approach could also result in loss of original information, however I believe that having a data set with such vast number of features, gives me freedom to experiment and train my models to an optimal extent.

The reason I believe that this dataset would do well with a Tree classifier such as Decision Tree is the fact that we know that these trees are effective for handling datasets with mixed feature types i.e. numerical and categorical as mentioned earlier. There are features that require my attention to be better formatted as well such as: ‘policy_bind_date’, ‘incident_date’.

I intend to find the optimal hyper parameters of the Decision Tree, while training in order to avoid model Overfitting or under fitting. Once I obtain the trained model I’d then proceed to compare important classification metric scores such as F-1, Precision, Recall and Accuracy against other classification models.

One challenging task would be to normalize this data and convert it completely into continuous data, by encoding categorical features and training the transformed data set on Logistic Regression and to see how it compares to the Decision Tree.

References

Federal Bureau of Investigation. Insurance fraud. Retrieved from https://www.fbi.gov/stats-services/publications/insurance-fraud

Insurance Information Institute. (2022, August 1). Background on: Insurance fraud. Retrieved from https://www.iii.org/article/background-on-insurance-fraud

Verisk Insurance Solutions. (2017). The challenge of auto insurance premium leakage: Mispriced risk is a $29 billion annual problem for the industry. Retrieved from https://www.verisk.com/siteassets/media/campaigns/gated/underwriting/verisk-the-challenge-of-auto-insurance-premium-leakage.pdf?__FormGuid=8c509869-699d-4698-9ac3-9ada3d271c97&__FormLanguage=en-US&__FormSubmissionId=65299f67-8c20-408c-97bd-221fc1cee1bf

Subudhi, S., & Panigrahi, S. (2017). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2017.09.010

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. Kollam, India. https://doi.org/10.1109/ICCPCT.2017.8074258

Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. *Applied Soft Computing*, 70, 1000-1009. https://doi.org/10.1016/j.asoc.2017.07.027

Hassan, A. K. I., & Abraham, A. (2016). Modeling Insurance Fraud Detection Using Ensemble Combining Classification. *International Journal of Computer Information Systems and Industrial Management Applications*, 8, 257-265. Retrieved from https://www.softcomputing.net/amira2016.pdf

ARPAN. (2022). Insurance Fraud Detection [Dataset]. Indian Institute of Management Calcutta & Kaggle. Retrieved from https://www.kaggle.com/arpan129/insurance-fraud-detection