Institute of Actuaries of India
Statutory body under an Act of Parliament

Expanding the horizon, strengthening the core
20th Global Conference of Actuaries
4th - 6th March, 2019 | Mumbai, India

Machine Learning models for Automobile Fraud Detection – A literature Review

Vamsidhar Ambatipudi, PGDM(IIM Indore), FIAI, CERA, FRM, PRM

Session # C5
Dated : 5-Mar-19

Agenda

- Introduction to Auto Insurance Fraud
- Impact of Fraud on Stakeholders
- Traditional Approaches to addressing Fraud
- Machine Learning Approaches
- The way Forward

Introduction

Fraud in Automobile Insurance

- Growing Competition in Automobile Insurance Industry
  - Frequency and Severity of claims increasing year on year
- Increasing Instances of Fraud in Auto Insurance (Moral Hazard)
- Fraudsters stage traffic accidents and issue fake insurance claims
- Fake reports of stolen vehicles
- Claims for pre-existing damages
- Soft vs. Hard Fraud
- Organized groups of fraudsters
  - Drivers, chiropractors, garage mechanics, lawyers, police officers, insurance workers etc.
- Fighting against insurance fraud is a challenging problem both technically and operationally

Introduction

Basic Statistics on Automobile Frauds

- An upward trend has been noted among many types of crimes, automobile insurance fraud included
    - Federal Bureau of Investigation Financial Crimes Report to the Public (2014)
- Approximately 20–35% of automobile insurance claims are fraudulent to a certain extent
- Statistics of the U.S. Coalition against Insurance Fraud illuminate that the amount of insurance fraud accounts for 17% to 20% of total insurance company compensation
- In China, insurance regulators estimate that the proportion of insurance fraud is approximately 20%
- In 2014, the Association of British Insurers (ABI) investigates the increase in number of false claims, which is 18% more than the previous year

Introduction

Common Features of Fraud Accidents

- Generally occur during the late hours and in suburban areas
- Limited scope of any witnesses
- Drivers are usually young males
- Accident cars are mostly private cars
- Police are always called to the scene (Make the subsequent claims easier and more reasonable)
- Many passengers in the vehicles, but never children or elders
- All of the participants have multiple (serious) injuries
- Almost no damage on the vehicles

Introduction

Factors that can influence Fraud

- Accident Information
  - Occurrence time and location, Reason for Accident, Time to Start and End, Repair shop type, Site Report, Investigation, Number of damage photos
- Insured Driver Information
  - Gender of policy-holder, Historical claims, Income and profession of the claimant
- Automobile Information
  - Type of automobile, Color of the vehicle, Brand, Registered Location
- Policy Related Information
  - Insurance maturity, Underwriting Type, Insured amount
- Text Description of the accident
- Certain correlation exists between selected indicators

Impact of Fraud on Stakeholders

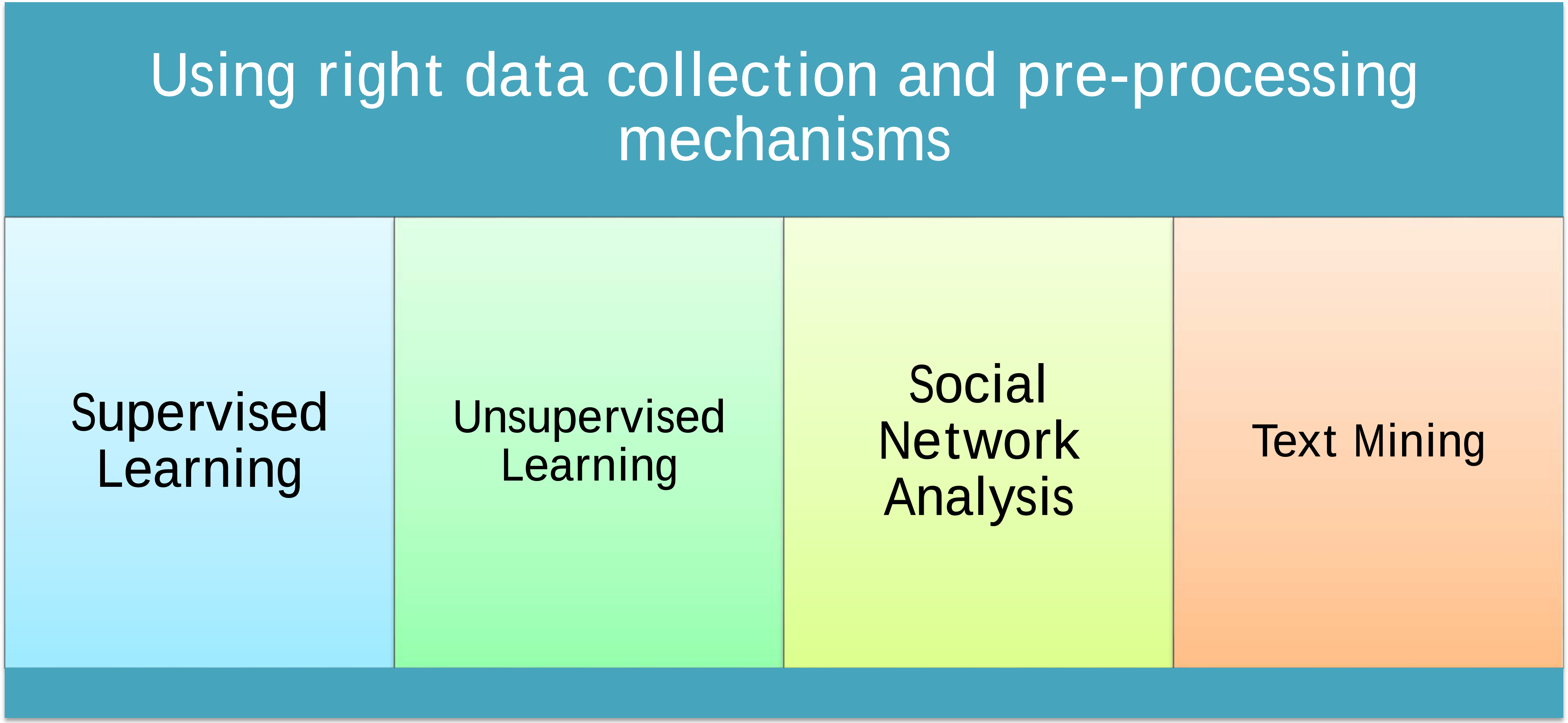
- Insurers
    - Lost Profits
    - Reputation losses
    - Increased cost of hiring highly skilled investigation team
- Policy Holders
    - Higher Premiums
    - Longer Settlement Times
    - Increased scrutiny on claims

Traditional Approaches to handle Fraud

- Auditing and Expert inspection
- Manual detection of fraud cases - costly and inefficient
  - Number of experts is negligible compared to the increasing number of claims
  - Average time that experts spend reviewing a case is not sufficient
- Rule based Approaches
  - Detect suspicious fraud cases in a timely manner (Scoring claims)
  - Relatively effective compared to that of experts in recognizing existing fraudulent claims
- Fraudsters are very innovative and new types of fraud emerge constantly (Adaptability is required)

Machine Learning Approaches

Overview of Different Methods



Using right data collection and pre-processing mechanisms

- Supervised Learning
- Unsupervised Learning
- Social Network Analysis
- Text Mining

Machine Learning Approaches

Supervised Learning

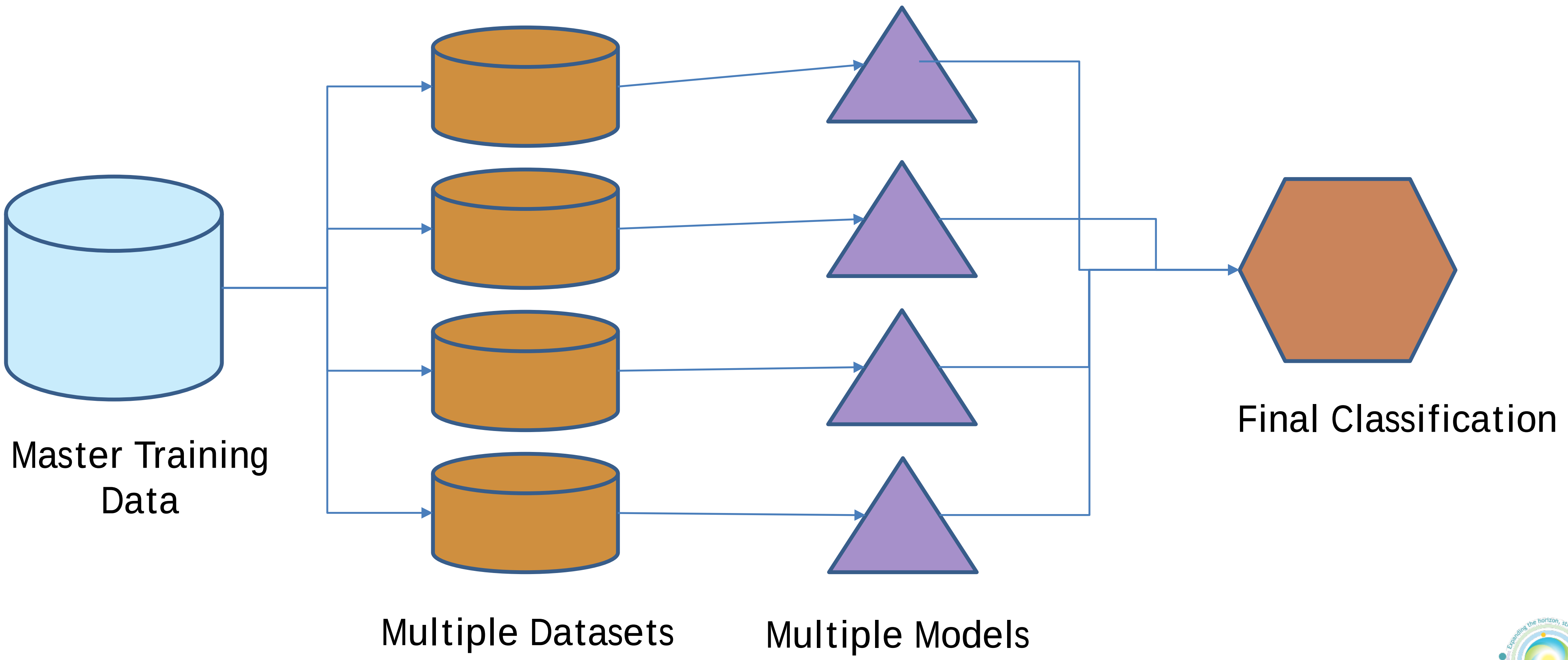
- Outcome variable (Fraud Indicator) is linked to explanatory variables
- Feature selection, Feature engineering to identify best predictors
- Relationships between the variables assessed
- Stand alone Models
  - Binary Choice models (Logit model and Probit model)
  - Decision Tree (DT) using CART
  - Support Vector Machine
  - Artificial Neural Network
  - Fuzzy Logic
  - Genetic Algorithms
- Weaknesses
  - Demand a labeled (initial) data set
  - Only suitable for larger, richer data sets

Supervised Learning – Ensemble Learning

- New machine-learning framework - Treats multiple learners as different modules to solve the same problem
- Improves the generalization ability of learning systems significantly
  - Boosting
  - Bagging
  - Random Decision Forests → Combines Bagging and Random subspace methods
    - Automatically selects features - Not sensitive to irrelevant features
    - Reduces possibility of overfitting
    - Considers the interaction between features
    - Applicable to binary class and multiple class problems
    - Does not require complex parameter selection process
    - Performance of RF depends on the classification ability of each tree and the diversity and correlation among them

Machine Learning Approaches

Supervised Learning – Ensemble Learning



Machine Learning Approaches

Supervised Learning – Advanced Approaches

- PCA – RF – PNN (Li et al., 2018)
  - PCA and retaining all components (Independent components but no loss of detail)
  - Potential Nearest Neighbours → Selects the samples that are nearest to the sample to be tested in the training set based on the principle of PNN
  - Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) to evaluate the importance of variables

Machine Learning Approaches

Supervised Learning – Advanced Approaches

- **Fuzzy C-Means Clustering and Supervised Classifiers (Subudhi & Panigrahi, 2017)**
  - Cluster Centers generated through Fuzzy C-Means clustering (FCM)
  - Genetic Algorithm (GA) for optimizing the cluster centers
  - A new insurance claim is classified as genuine, malicious or suspicious based on its distance computed from the optimized cluster centers
  - Suspicious claims are identified among the insurance records and their behavior is further verified by supervised classifiers
    - SVM, Multilayer Perceptron (MLP), GMDH (Group method for data handling) and Decision Tree

Machine Learning Approaches

Unsupervised Learning

- Obtaining labels is costly and time consuming
- Clustering
  - Frauds and Non-frauds can be separated based on distance between the attributes
  - Traditional clusters as well as Neural Network based clusters exist
- Distance based identification of outliers and thus frauds
  - KNN Based, Density based methods
- Spectral Ranking for Anomaly Detection (Nian et al., 2016)
  - Ranking represents the degree of relative abnormality

Machine Learning Approaches

Social Network Analysis (Subelj et al.,2011)

- Helps in Detecting Organized collusion activities
  - Detection of groups of collaborating fraudsters, and their connecting accidents
- Combine intrinsic attributes of entities with their relational attributes (based on graphs)
- Iterative Assessment Algorithm
  - Allows detection of new types of fraud as soon as they are encountered (No learning from labelled data)
- Uses networks of collisions to assign suspicion score to each entity

Machine Learning Approaches

Social Network Analysis - Steps

- Different types of networks are constructed from the data set
    - Driver Networks, Participants Networks, Connect Passengers Through Accidents networks, Vehicle Networks
- Components are investigated based on their structural properties(diameter, cycles etc.) – Suspicious ones are identified
    - Fraudulent components share several structural characteristics (much larger and Denser)
    - There are vertices with extremely high degree and centrality
- Suspicious Components are further analyzed in order to detect key entities inside them (assigns a suspicion score to each entity)
- Obtained results should always be investigated by the domain expert or investigator

Machine Learning Approaches

Text Mining

- Mining of different documents collected during investigation
  - Topic Modelling/Sentiment Analysis
  - Clustering & Other Pattern identification
  - Combining Classification and Topic Models
- Latent Dirichlet Analysis (LDA) based Text Analytics (Wang & Xu, 2018)
  - Combines the precision of data mining methods and the experience of human experts
  - Generative model that can be used to identify hidden topic information in large-scale document collections
  - Obtain the distribution of the keywords for each topic and the distribution of the topics in each description
  - Deep learning is employed to seek high-quality attributes
  - 10-fold cross-validation is used to ensure the validity

Machine Learning Approaches

Technical Difficulties in Fraud Detection

- Skewed Class Distribution → only a small portion of accidents or participants is fraudulent
  - Collaborated effort of all insurers to develop a master fraud database
- Lack of labeled data sets → labeling is expensive and time consuming
  - Joint Knowledge sharing based on experience of new frauds
- Lack of unlabeled data sets
- Garbage In Garbage Out
  - Data capturing mechanisms need to be improved (Integrated and verified with various sources of information)

References

- 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.
- Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58-75.
- 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.
- Šubelj, L., Furlan, Š., & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems with Applications, 38(1), 1039-1052.
- Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 105, 87-95.

Expanding the horizon, strengthening the core

20th Global Conference of Actuaries

4th - 6th March, 2019 | Mumbai, India

THANK YOU

Analytics for Better Solutions