According to the coalition against Insurance Fraud, 10% of the total claims filed for P&C insurance are estimated to be fraudulent, and conservatively Fraud (all lines of Insurance) costs 80 billion dollars annually.

ABSTRACT

The impact of fraud related losses is increasingly severe as new and innovative ways to defraud insurance companies are on the rise. While all insurance companies have a mechanism to detect insurance fraud, most of them do so using a traditional approach, based on business knowledge of insurance SMEs and SIUs, by creating a rule based fraud detection engine. However, a new trend of using data analytics for prediction of insurance fraud has emerged over the last few years.

This paper takes a look at this emerging trend and how it is being used by insurance companies to predict fraudulent claims in the auto insurance line of business. It takes a deeper look at the various analytical techniques that enable accurate fraud prediction to help Insurance companies to optimally utilize resources and reduce fraud losses.

BACKGROUND

Fraud in Insurance has existed from the beginning of the industry itself. In recent times, there has been an increase in the incidence of insurance fraud being committed. A large percentage of the consumers think it is acceptable to defraud their insurance company. Also instances of insurance fraud being committed by organized rings are on the rise, as they view insurance fraud as less dangerous.
Both the insurer and the insured are impacted by insurance fraud either directly or indirectly. For the insured party:

- The premiums go up as Insurers cover their losses through increased premiums from the customers
- There are increased scrutiny measures and longer periods to settle claims

For the Insurance Companies, the main impact is felt as:

- Lower profitability as any money spent on fraud impacts the bottom line
- Higher human capital cost of employing Fraud Investigation Units
- Frauds when not managed effectively could cause loss of business, as premiums increase to compensate for the losses and may no longer be competitive.

Traditional techniques are not sufficient to combat new and innovative methods used by fraudsters and to effectively manage insurance fraud. Advanced analytical techniques enable insurance companies to mine their data, identify emerging patterns and create predictive models to predict fraudulent transactions. In addition to this, collaboration between insurance companies and external data agencies has also become a key weapon for companies to combat insurance fraud.

Accurate predictions of fraudulent insurance claims result in:

- Minimizing fraud loss, and thus improving the bottom line for insurance companies
- Straight through processing of clean claim transactions, thus improving the operational efficiency of claims adjusters and overall reduction in claim-processing costs

www.insurancefraud.org
www.iii.org/issue-update/insurance-fraud
ADVANCED ANALYTICS FOR FRAUD MINING

The know-how of the insurance Subject Matter Expert (SME) and internal and external data are key assets in fighting against fraud. The rule based approach and data analytics approach utilize these assets to discover fraudulent transactions in claims data. It is important to use a combination of these approaches to bring out the best strategy to discover fraud.

The rule based approach, identifies red flag variables and rules. Using these rules, it scores the claims, which provides an indication of a claim being fraud or not.

In data analytics approach, hidden patterns in the data are discovered, which helps us to accurately predict fraud.

Listed below are the key data analytics techniques used for this purpose. All of these techniques together will help create a robust fraud prediction solution.

1. Unsupervised Learning Techniques:

Insurance analysts believe that 10% of all auto insurance claims are fraudulent, but only a part of them are caught through traditional methods. Insurance companies investigate suspicious claims, but cannot investigate all of them due to the costs attached. Hence, many fraudulent claims slip through and are undetected. If one has to build an effective predictive model, the data required for building the predictive model should have the fraud indicator variable set correctly for all or most of the claims in historical data. If that is not the case, the predictive model will target to identify only those patterns for which the fraud indicator variable is set to “Yes”, as fraudulent. Hence, to provide correct data input to the predictive model, all historical non-investigated data should also be classified correctly into suspicious and non-suspicious claim transactions.

This is achieved through using unsupervised learning models to identify the anomalies or outliers in the data which are indicative of suspicious claim transactions. Normally, these findings are reviewed and validated by Insurance SMEs to improve the model and increase confidence in the model. Apart from this, unsupervised models can also be used to identify anomalies in the live data and score the live data.
2. Predictive Analytics:

Predictive Analytics are used to build analytical models which use historical data (where the value of the outcome variable is known) to build a model, which can predict the value of the outcome variable in new data where that value is not known. A good predictive model can accurately predict the value of the outcome variable and thus help with quick decisions in the process workflow.

Predictive analytics uses an outcome variable, which, in the fraud prediction case, is the fraud indicator variable, for building the predictive model. As a first step, it is important to understand what data is useful and available for building the fraud model. Normally, the historical claim, policy, customer profile data, investigation data, along with the classified and reviewed, non-investigated data are used for this purpose. Allied data points such as driving license and ticket-related can also be useful.

The predictive model building process has to go through several steps such as assessing the quality of data, understanding the variables and relationships between them, selecting the best predictor variables and model building and validation.
There are several statistical and machine learning algorithms used for classification to predict insurance fraud. These include -

- Decision Trees
- Support Vector Machines
- Neural Networks
- Boosting
- Bagging
- Random Forests

While selecting a particular model for making predictions, it is important to take into consideration the characteristics of the data, the strengths and weaknesses of the algorithm, accuracy of the algorithm to make correct predictions, and speed and scalability of the algorithm. General pitfalls to avoid are

a. Avoid using meaningless or noise variables for model building
b. The class imbalance problem should not be overlooked
c. Over fitting the model to training data should be avoided

Accuracy of the model can be viewed through the confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>Predicted Not Fraud</th>
<th>Predicted Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual, Not Fraud</td>
<td>Correctly predicted, faster claim processing</td>
<td>Results into unnecessary investigation, payout delays</td>
</tr>
<tr>
<td>Actual Fraud</td>
<td>Results in fraud Loss</td>
<td>Correctly predicted, prevention of fraud loss</td>
</tr>
</tbody>
</table>

For many of the algorithms, like Random Forest, CART, CHAID, it is possible to add a penalty for wrong prediction. E.g. if avoiding fraud loss is more important than avoiding unnecessary investigation for non-fraud cases, one can add a penalty for incorrect prediction of a case as not fraud. This will reduce “Incorrect prediction as Not Fraud” and improve accuracy of correct fraud prediction.

Ensemble models are widely used for insurance fraud prediction. Ensemble models are based on the idea that by combining weak learners, a strong learner is created. Bagging, boosting and random forest are examples of ensemble models. Advantages of using ensembles are

a. Reduces chance of over fitting – No single bias is able to dominate because opinions of several learners are incorporated.
b. Improves performance – Since ensembles create several smaller models as against a single large model, computing performance can be improved. Also, it is easier to parallelize an ensemble using distributed computing methods
c. Provides a better understanding of difficult learning tasks – Real world problems are complex. Models that divide the task into smaller parts can capture subtle patterns more accurately than a single model might.
d. An ensemble like random forest can be used with data with extremely large features or examples. It selects most important features and performs well on most of the problems.
3. **Social Network Analysis**

It is possible that several people are colluding together to defraud an insurance company. They may be known to each other or can also be part of a fraud ring. This aspect is missed by most of the traditional fraud detection methods. Frauds which are committed by such a coalition may not appear as a fraud when looked at individually. But when these are looked at as a group, it is possible to see common patterns between their claims. Social network analysis brings out such patterns in the data and helps to discover people/networks committing these frauds.

Algorithms such as social network analysis, Google page rank, association algorithms are used to bring out these insights from the data. Post this investigations into those networks need to happen to confirm the presence of fraud rings.

4. **Text Analytics**

Text Analytics is also a new technique available for fraud prediction. The claim settlement process creates several documents. Some of these documents, such as a police report, medical report, adjustor notes, etc., may provide an indication of fraudulent transactions. Most of the time, these reports can be paper based and stored in a document management system as a scanned pdf or image files. These reports can be a rich source of data to predict fraud. So using an optical character recognition tool, one can convert these scanned PDFs/images to structured and unstructured text data. Text analytics techniques are used on unstructured text to convert it into insights.

Text analytics and sentiment analytics can be used on unstructured text data to identify feelings, attitudes and opinions. So by analyzing the doctor’s notes, adjuster’s notes, police reports, etc. opinion can be mined and used for analysis.

Text Analytics can also be used for indexing the unstructured notes and using the additional variables for building predictive models.
Tech Mahindra has created a solution framework, integrated Fraud Analytics Solution (iFAS), for detection of the insurance frauds. iFAS provides an end to end framework from prediction to investigation of auto insurance fraud claims.

The key features of iFAS include:
- Uses a hybrid rule and data analytics based scoring approach
- Uses optical character recognition for conversion of scanned PDFs/images to structure and unstructured text
- Uses advanced analytics techniques such as anomaly detection, predictive modeling, social network analysis and text analytics for prediction of fraudulent claims
- Real time or batch scoring options for claim scoring
- Seamless integration with claim management systems and document management systems
- Easy integration with internal and external data sources
- Modular design and incremental delivery approach

KEY TAKEAWAYS

The Data Analytics approach provides a new approach for prediction of fraud using the organizations data as an asset. More and more insurance companies have started to incorporate this approach in their fight against fraud. There are several techniques such as predictive modeling, text analytics, unsupervised learning, and social network analysis that help to bring out insights from historical data and use them for prediction of frauds in new claim data. For each of these techniques, there are multiple algorithms available which allow organizations to take advantage of these techniques. These techniques and algorithms for these techniques have provided new tools to insurance organizations to combat fraud.
Experience:
Having over 18 years of experience in IT industry spread across Banking, Financial services and Insurance domains in Big data, Analytics, Business Intelligence and Information Management technologies.
Currently focusing on Big data and analytics solutions for Insurance and other verticals
Hands on experience in designing and consulting on BI, Big data and analytics solutions. Deep expertise and experience in Insurance domain.

PRASHANT GHATAGE
Principal Consultant - Big Data and Analytics

Education:
B Tech from IIT Powai