Paper 1837-2014

The Use of Analytics for Claim Fraud Detection

Roosevelt C. Mosley, Jr., FCAS, MAAA
Nick Kucera
Pinnacle Actuarial Resources Inc., Bloomington, IL

ABSTRACT

As it has been widely reported in the insurance trade news, fraudulent claims continue to be a significant issue in the insurance industry, costing policyholders billions of dollars. More companies are turning to analytics to help identifying claim fraud. Identifying claim fraud using predictive analytics, however, represents a unique challenge.

1. Most predictive analytics applications have a complete target variable which can be analyzed. Fraud is unique in that there is generally a lot of fraud that has occurred historically that has not been identified. Therefore, the definition of the target variable is not complete.

2. There is a natural assumption that the past will bear some resemblance to the future. In the case of fraud, methods of defrauding insurance companies change quickly and it can make the analysis of a historical database less valuable for identifying future fraud.

3. In an underlying database of claims that may have been determined to be fraudulent by an insurance company, there are inconsistencies between different claim adjusters regarding which claims are referred for further investigation. These inconsistencies can lead to historical databases that are not complete.

This paper will demonstrate how analytics can be used to help identify fraud and allow an insurer to optimize the resources they have available in combating this fraud. Applications discussed include:

1. More consistent referral of suspicious claims to claim investigative units

2. Better identification of suspicious claims, even as techniques used to defraud insurers are changing

3. Incorporating claim adjuster insight into analytics results to improve the process

As part of this paper, we will demonstrate the application of several approaches to fraud identification:

1. Clustering

2. Association analysis

3. PRIDIT (Principal Component Analysis of RIDIT scores)

INTRODUCTION

In the property and casualty insurance industry, there are many reports that the occurrence of claim fraud is increasing, and it is evident that the focus on claim fraud has been magnified. Many of the estimates of the amount of claim fraud in the industry show this trend, and one only has to follow an insurance news feed to know that the amount of reporting that is related to claim fraud seems to be rising significantly. Regardless of the exact amount of fraud present in property and casualty insurance, all agree that it is a significant amount and thus a concern that needs to be addressed.

Insurance companies have developed effective procedures for identifying, investigating, and deterring fraudulent activity. The combination of experienced claim adjusters and special investigators has produced a process that ensures claim payments being made are fair and legitimate. The experience, insight, and intuition of these claims personnel have saved insurance companies millions of dollars in payments over the years.

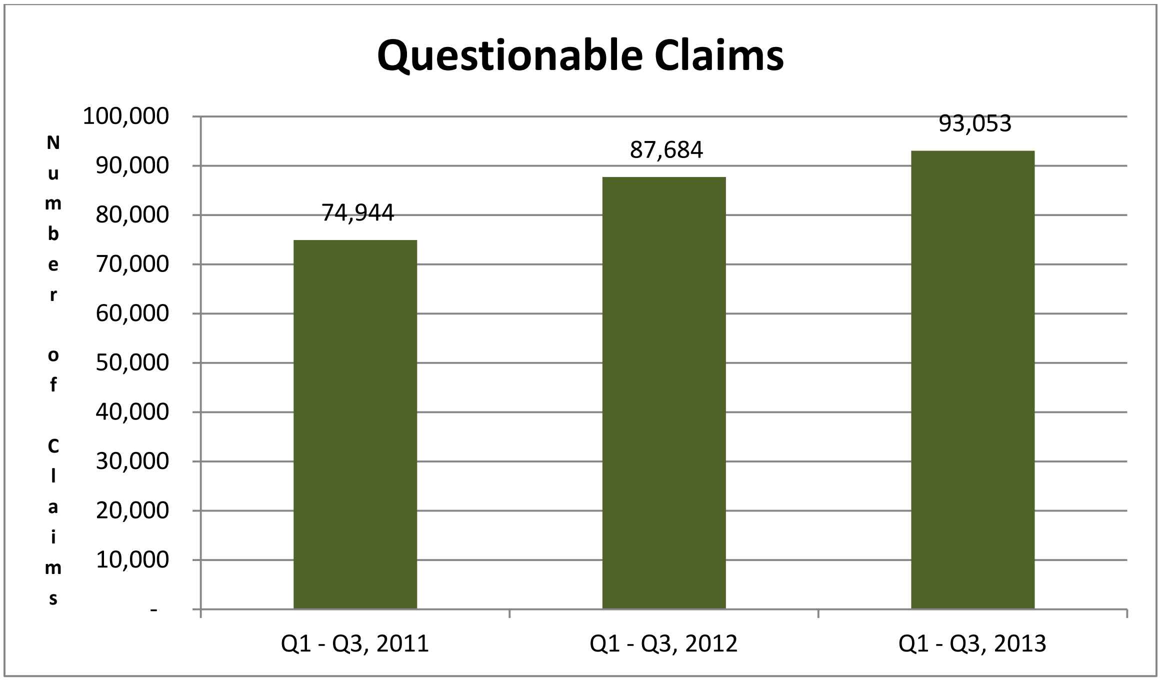
However, as good as experienced claim adjusters and special investigators are, the reality is there are not enough of these trained eyes to review every claim, and thus there are some fraudulent claims that slip through the cracks. As a result, payments are sometimes made that should not be.

Predictive analytics can assist insurance companies in developing a more consistent claim referral process, such that the benefit of the expertise of the best adjusters and investigators is applied to all claims. Predictive analytics can also enhance the work of the claims department by uncovering complexities and nuances in a particular claim that may be missed by even the most experienced claim adjusters.

THE CLAIM FRAUD PROBLEM

The problem of claim fraud by any measure is a large one. This can be seen in fraud statistics that are tracked by different organizations and the financial impact it has on the industry.

The National Insurance Crime Bureau (NICB) produces a quarterly report on questionable claims reported to NICB by insurance companies. Based on the latest report, the number of questionable claims for the first three quarters of 2011 as compared to 2009 and 2010 are shown below[1].



[1] Fennig, David. "First 3 Quarters of 2011, 2012, 2013 Questionable Claim Referral Reason Analysis (Public Dissemination)." NICB Data Analytics Forecast Report. October 29, 2013.

As can be seen from the chart above, the number of questionable claims has increased significantly in 2012 (17.0%) and 2013 (6.1%).

The Coalition Against Insurance Fraud also conducted a study of consumer attitudes related to insurance fraud. The problem of fraud is highlighted in some of the key findings from this study[2].

- 1 in 5 adults think it is acceptable to defraud insurance companies
- 1 in 10 people think it is OK to submit claims for items that are not lost or damaged, or for an injury that didn't occur
- 16% of people think it is OK to inflate a claim to cover the deductible

These types of consumer attitudes compound the problem, and are further confirmation that fraud is a significant issue for insurance companies.

Fraud costs insurance companies, and ultimately consumers, a significant amount of money. The Coalition Against Insurance Fraud estimates that fraud and buildup added \$4.8B to \$6.8B to auto insurance costs in 2007[3]. In worker's compensation, fraud cost insurers over \$550 million dollars in 2008[4]. In healthcare, \$68B is lost to fraud every year[5]. In total, the Coalition estimates that \$80B is lost due to fraudulent claims every year. With billions of dollars at stake, even a minimal increase in fraud detection or prevention will translate into substantial dollar savings for insurance companies and their customers.

THE FRAUD DETECTION PROCESS

Given the significant cost, companies have developed processes to combat insurance fraud. A typical fraud detection process is shown below:

[2] Coalition Against Insurance Fraud. "Four Faces: Why Some Americans Do – And Don't – Tolerate Fraud." October, 1997.

[3] http://www.insurancefraud.org/autoinsurance.htm

[4] http://www.insurancefraud.org/workerscompensation.htm

[5] http://www.insurancefraud.org/healthinsurance.htm

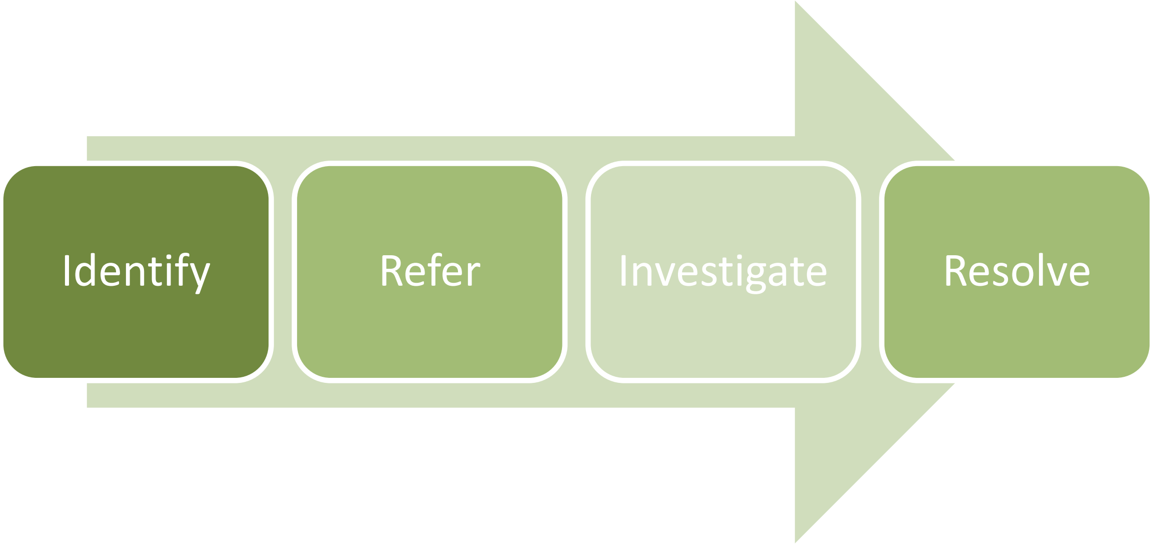


Figure 1: Insurance Company Fraud Detection Process

The first step in the fraud detection process is the identification of a suspicious claim. Historically, there have been two main ways that these claims have been identified. The first has been to rely on adjusters to identify potentially fraudulent claims. The second has been by using a set of fraud indicators to assist the claim adjuster in determining if the claim is suspicious and warrants additional scrutiny.

One way an adjuster identifies a suspicious claim is by recognition of a suspicious pattern that has been seen previously. An example of this could be a repeat offender, or if the same medical provider, attorney, and patient show up together repeatedly in different claims. This ability is heavily dependent on the experience of the claim adjuster, and also on access to advisory claim databases where information on historical claims can be researched and the current claim validated in light of this historical information. This approach to identifying suspicious claims has some disadvantages though. This assumes that the investigator has seen this type of fraud in the past, so it will be heavily dependent on the experience level of the adjuster. Also, as fraudsters adapt and become smarter (i.e. using aliases, including different individuals in a fraudulent network), repeat patterns will be more difficult to recognize.

Another approach that adjusters use to identify suspicious claims is simply applying their experience and intuition. Generally, if something “smells funny” to a seasoned adjuster about a claim, it can be a great indicator that there is a potential issue with the claim. It is at this point that the claim is then referred to the special investigative unit (SIU). This approach relies as well on the experience of the adjuster to identify suspicious claims. Because of this, the obvious drawback is that adjusters with less experience may not be able to detect that the claims are suspicious.

To address the issue of consistency, some companies supplement the experience of the adjuster with fraud indicators. These are rules developed by the company that determine if a claim should be referred for further investigation. This approach can be useful in identifying known and typical fraud scenarios. There are a few advantages to an approach like this. These advantages include the facts that it is easy to implement and modify, it is easy to understand, it is effective in attacking specific problems, and when implemented, produces consistent results. However, there are disadvantages to this approach as well. First, it does not help in detecting new and unknown fraud scenarios. It also has the additional problem of creating smarter fraudsters as they attempt to circumvent the fraud indicators.

Example of fraud indicators are:

- Distance between claimant’s home address and medical provider
- Multiple medical opinions/providers
- Certain injury types (e.g., soft tissue)
- Changing providers for the same treatment
- Higher than average number of treatments
- Abnormally long time off for a given type of injury
- Loss payments that do not correlate with the severity of the injury

Based on these fraud detection methods, suspicious claims are then referred to the SIU for further investigation. However, the issues with the claim identification methods create challenges for the SIU.

1. The claims which are referred for further investigation can be inconsistent, which is going to be heavily dependent on the adjuster that is handling the claim.

2. The claim identification process can produce a significant number of false positives. False positives result in less than efficient use of investigation resources.

3. Claim adjusters may not be aware of all the potential suspicious relationships, and therefore may miss some claims that really should be investigated further.

4. Because not all fraudulent claims have been identified historically, it is difficult to identify these missed patterns going forward.

5. There have been cases where a claim adjuster has been involved in the scheme to defraud an insurer. Obviously, in a case like this relying on an adjuster to identify fraud will not work.

6. After suspicious claims are identified, it can be very difficult to prioritize these claims. Prioritization is an important issue since most SIU’s are not adequately staffed to investigate every suspicious claim identified.

Predictive analytics can be used to address the concerns with historical claim fraud processes and to optimize the resources a company has at its disposal.

USING PREDICTIVE ANALYTICS TO DETECT SUSPICIOUS CLAIMS

There are multiple ways in which predictive analytics can be used to detect suspicious claims.

1. Analysis of Historical Referrals

2. Analysis of Historical Fraudulent Claims

3. Identification of Networks

4. Identification of Suspicious Claim Patterns

5. Combination of Analytics and Adjuster Experience

Analysis of Historical Referrals

This first analysis allows insurance companies to apply analytics for more consistent claim referrals to the SIU. For this analysis, the target variable is a history of claim referrals to the special investigative unit. This target can be adjusted to just use the referrals of those claim adjusters that are considered the most experienced or most effective. The independent variables considered in the analysis are the details of the claims. The result of the analysis is, based on the historical claim referrals, the likelihood that a new claim should be referred to the investigative unit.

The chart below shows the partial results of a decision tree based on an analysis of a database of automobile insurance bodily injury coverage closed claims. The dataset included an indicator if the claim was referred to an investigative unit, and also included the details of the claim (injury details, accident severity, claimant characteristics, location, and payment information).

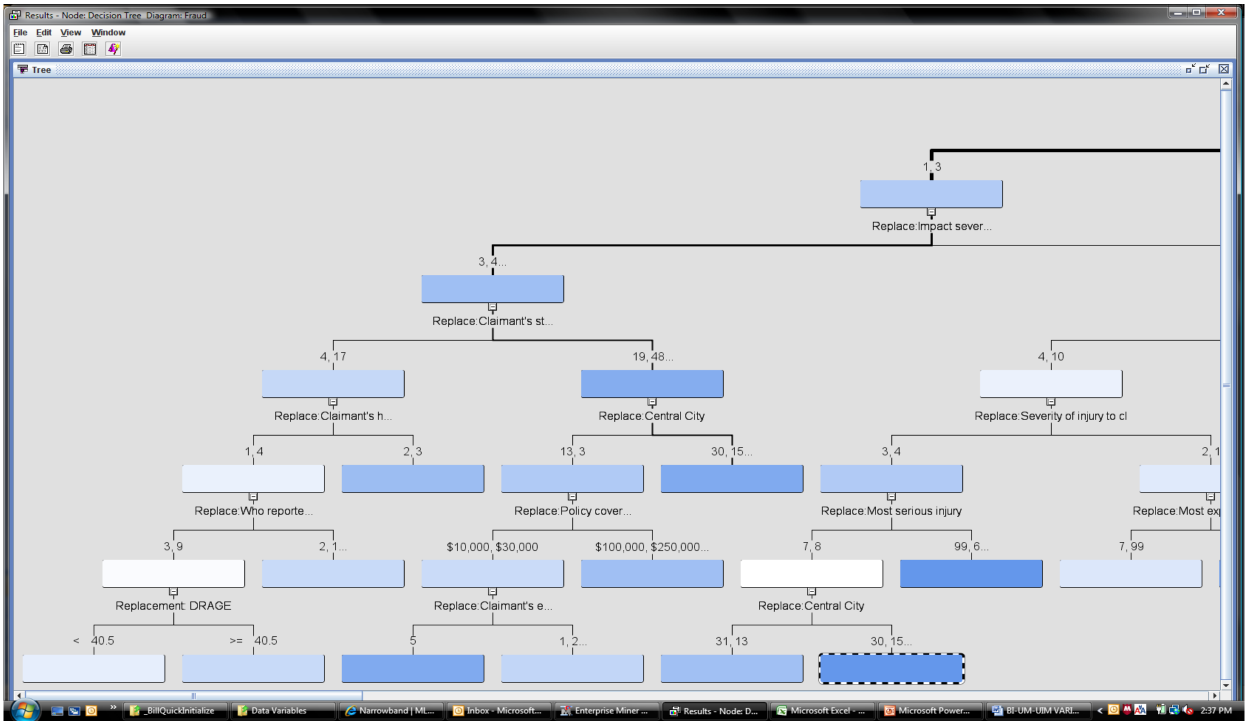


Figure 1: Decision Tree - Predicted Claim Referral

It can be determined from this analysis which factors were associated with increased claim referrals. These characteristics included the severity of the injury, the location of the accident, and whether the claimant was represented by an attorney. While the decision tree results are shown here, three separate predictive models were developed: Decision Tree, Neural Network, and Logistic Regression. Ultimately, the predicted referral was determined by an Ensemble model which used the combined effect of the three models to produce a more accurate prediction.

The distribution of the predicted likelihood of referral for the claims in the bodily injury dataset is shown below.

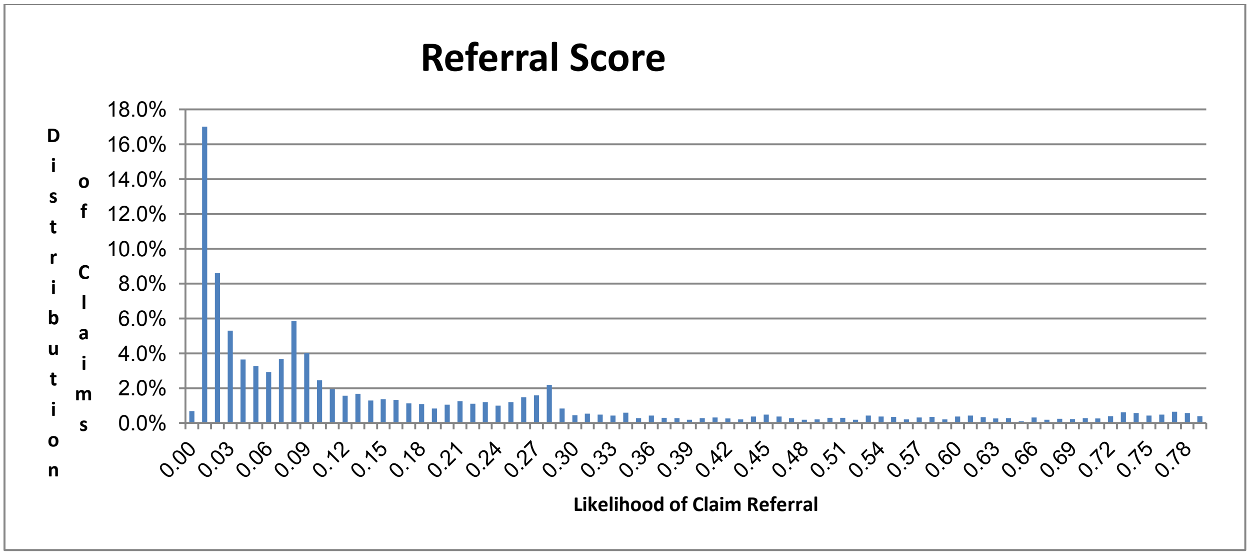


Figure 2: Predicted Likelihood of Claim Referral

Based on this distribution, a cutoff score could be chosen such that claims above a certain point would automatically be referred to the investigative unit. These claims could be sent directly to the SIU, allowing experienced claim adjusters to spend their time on those claims the model classifies as being borderline in regards to fraud likelihood.

<u>Analysis of Historical Fraudulent Claims</u>

The analysis of historically fraudulent claims relates to claims that were not just referred to an investigative unit, but those claims on which action was taken. This action can include a lower claim payment, the denial of a claim, and/or the referral of a claim to law enforcement authorities. So the target is whether action was taken on the claim, and the independent variables are the same as those used for the referral analysis.

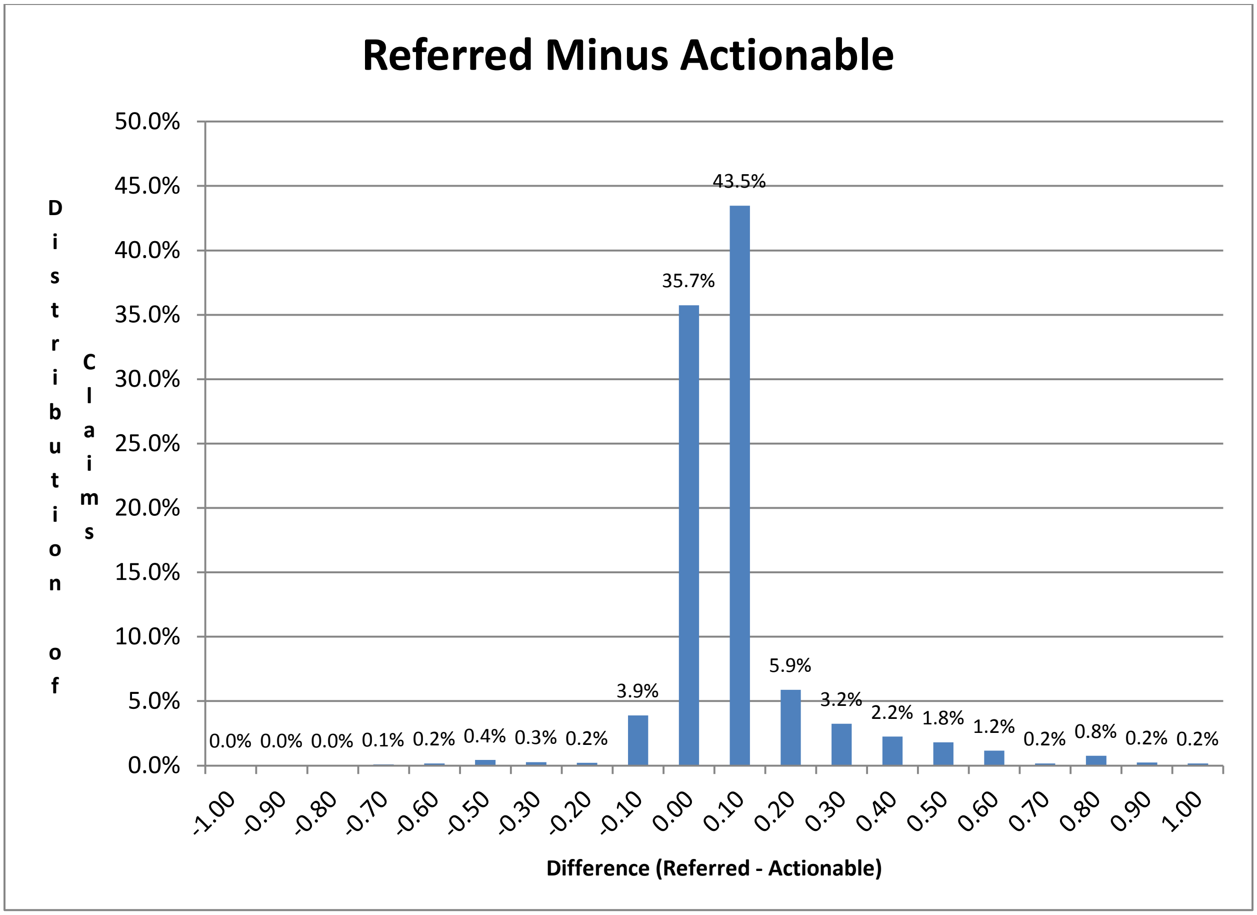
From the same database, we analyzed the likelihood of action being taken on the claim. Not only did we review the predicted likelihood of action, but we also compared the results of the predicted fraud model to the predicted referral model. This allows insurers to understand the difference in the variables that drive referrals vs. action, and also to refine the referral process to reduce false positives and increase the focus on the true fraudulent claims that may have been slipping through the cracks.

The table below shows a comparison of the variable importance statistics from the decision trees from the actionable analysis and the referral analysis.

Variable	Actionable Importance	SIU Referral Importance	Ratio
Central City	1.000	0.464	46.4%
Replace:Claimant's state of residence	0.967	1.000	103.5%
Impact severity to claimant's vehicle	0.962	0.828	86.2%
Was claimant represented by an attorney?	0.850	0.905	106.4%
Policy coverage limits per person	0.750	0.411	54.9%
Arbitration	0.547	0.368	67.2%
Most serious injury	0.530	0.375	70.9%
Who reported injury to insurer	0.439	0.374	85.3%
Most expensive injury	0.423	0.239	56.5%
DRAGE	0.312	0.306	98.0%
Lawsuit status	0.295	0.000	0.0%
Driver, other violation	0.285	0.000	0.0%
Amount Spent on Medical Professionals	0.255	0.412	161.6%

In this table, it can be seen that for many of the variables, the importance is consistent between the two analyses. However, there are some of the variables where the importance is not consistent. For example, where the accident occurred (central city) is the most significant variable in the actionable analysis, however it is significantly less important in the referral analysis. Therefore, this may be an area that needs to be given more emphasis in the claim department. There are also variables that were important in the actionable analysis, but showed no importance in the referral analysis (lawsuit status, driver violation). This may point to areas that the claim department needs to consider in their referral process.

The chart below shows the difference between the predicted referral likelihood and the predicted actionable likelihood.



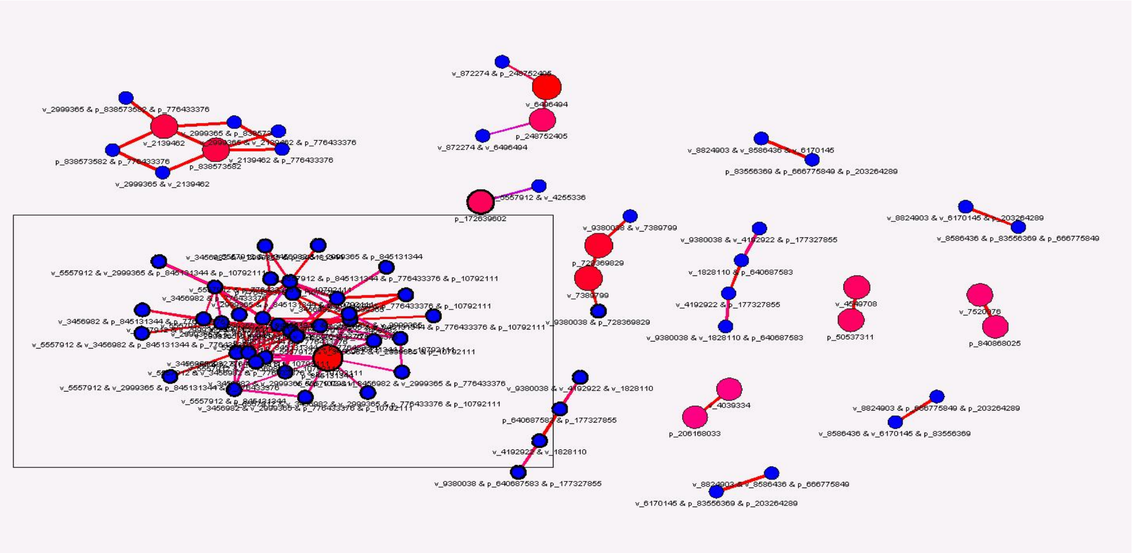
As can be seen from the chart above, the predicted referral likelihood and the predicted actionable likelihood are consistent for the majority of the claims. However, there are claims on both ends of the scale that should be investigated further. On the left side of the chart are claims with a predicted actionable likelihood that is significantly higher than the predicted referral likelihood. This represents claims that were not likely to be referred historically but should have been. The right side of the chart represents claims with a referral likelihood significantly higher that the actionable likelihood. These claims represent false positives, those claims that have been referred but turned out to be legitimate. Investigation of these claims drains a company's resources and represents an opportunity cost to the company's SIU. The results of this analysis can be used to refine the claim referral process and increase overall efficiency and effectiveness.

Association Analysis

Claim fraud can also be identified by determining suspicious associations in a claim analysis. There are a number of people associated with the claims process, including the insured, the claim adjuster (either internal or external), and any claim service provider (medical provider, auto repair shop, home repair contractors, etc.). There may be certain networks that show up consistently that raise suspicions and thus need to be investigated further.

This can be analyzed in SAS® Enterprise Miner using the Association Analysis. Association Analysis has its background in market basket analysis. It is used in retail environments, such as grocery stores or pharmacies, to identify items that tend to be purchased together. This analysis determines the likelihood of a combination of items occurring together as well as a confidence around the projection. Ultimately, the association analysis produces a set of if-then rules (if item A is present in a transaction, then item B will be present as well), and the lift associated with the rule. Applied to claim analyses, the items would represent those involved in the claim process, and the claim would be equivalent to the transaction.

The chart below shows the Association Map that results from an analysis of the claimants, claim adjusters, and repair providers for a homeowners claim analysis.


v_872274 & p_248752405
v_6496494
p_248752405
v_872274 & v_6496494
v_2999365 & p_838573582 & p_776433376
v_2139462
v_2999365 & v_2139462 & p_776433376
p_838573582 & p_776433376
p_838573582
v_2139462 & p_776433376
v_2999365 & v_2139462
v_8824903 & v_8586436 & v_6170145
p_83556369 & p_666775849 & p_203264289
v_5557912 & v_4255336
p_172639602
v_9380038 & v_7389799
p_728369829
v_7389799
v_9380038 & p_728369829
v_9380038 & v_4192922 & p_177327855
v_1828110 & p_640687583
v_4192922 & p_177327855
v_9380038 & v_1828110 & p_640687583
v_8824903 & v_6170145 & p_203264289
v_8586436 & p_83556369 & p_666775849
v_4549708
p_50537311
v_7520976
p_840868025
v_2999365 & p_845131344
v_5557912 & v_2999365 & p_845131344 & p_10792111
v_5557912 & p_845131344 & p_776433376 & p_10792111
v_3456982 & p_776433376
p_776433376 & p_10792111
v_3456982 & p_845131344
v_5557912 & v_3456982 & p_845131344 & p_776433376
v_5557912 & v_2999365 & p_845131344 & p_776433376
v_5557912 & p_845131344
v_3456982 & v_2999365 & p_776433376
v_3456982 & v_2999365 & p_776433376 & p_10792111
v_9380038 & v_4192922 & v_1828110
p_640687583 & p_177327855
v_4192922 & v_1828110
v_9380038 & p_640687583 & p_177327855
v_4039334
p_206168033
v_8824903 & p_666775849 & p_203264289
v_8586436 & v_6170145 & p_83556369
v_8824903 & v_8586436 & p_666775849
v_6170145 & p_83556369 & p_203264289

In the bottom left corner, there are a series of inter-connected participants which are all related to one repair provider (shown by the bright red circle). There are also a series of other connections that are outside of the main set of connections in the bottom left corner. It may be legitimate that some of these connections are there, as certain adjusters may be familiar with certain contractors, or certain providers are preferred for certain types of claims. It is at this point of the analysis that an experienced adjuster should review the results to ensure that there are no suspicious connections. Examples of things to look for would include historical experience with the contractor or past instances of known fraud.

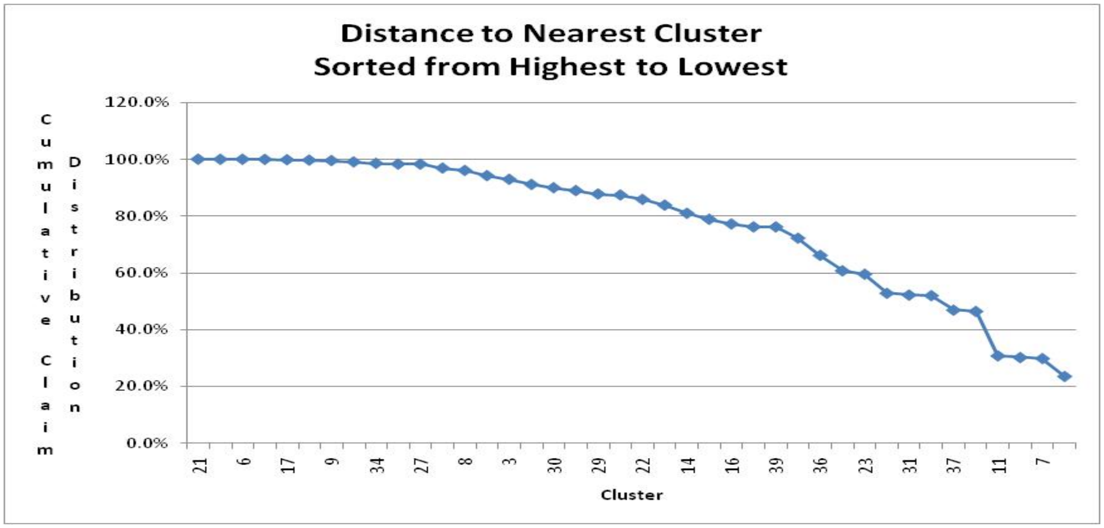
Identification of Suspicious Claims

The development of predictive models for referrals and actionable claims is important because it ensures that the company is protecting itself against known and identified fraud. However, this is a lagging indicator, because in order for certain types of fraud to be picked up by these predictive models, it needs to have been identified historically. Herein lies the problem with using a traditional predictive analytics structure, the target is not complete. This is due to the fact that all fraud that has been present in claims historically has not been identified. Also, new fraud adapts and will not necessarily be picked up by the historical models.

An approach that can be used to identify suspicious claims without a historical target is the use of clustering analyses. In a clustering analysis, claims are categorized into groups based on their characteristics. The more similar two claims are, the more likely they are to be grouped together. The more dissimilar the claims, the more likely they are to be classified into separate groups.

This approach yields three ways to identify suspicious claims. The first approach is to review the overall cluster statistics to determine if the cluster as a whole is suspicious. This is done by looking at how far the cluster is from other clusters, measured by looking at the distance to the nearest cluster. If it is far away from other clusters, as a whole it contains claims that can be considered outliers. Upon further investigation of the claims, it can be determined if the cluster description appears to describe claims that are suspicious.

The table below shows the results of a cluster analysis on homeowner content claims. The distance to the nearest cluster is charted, with the clusters sorted by decreasing distances.



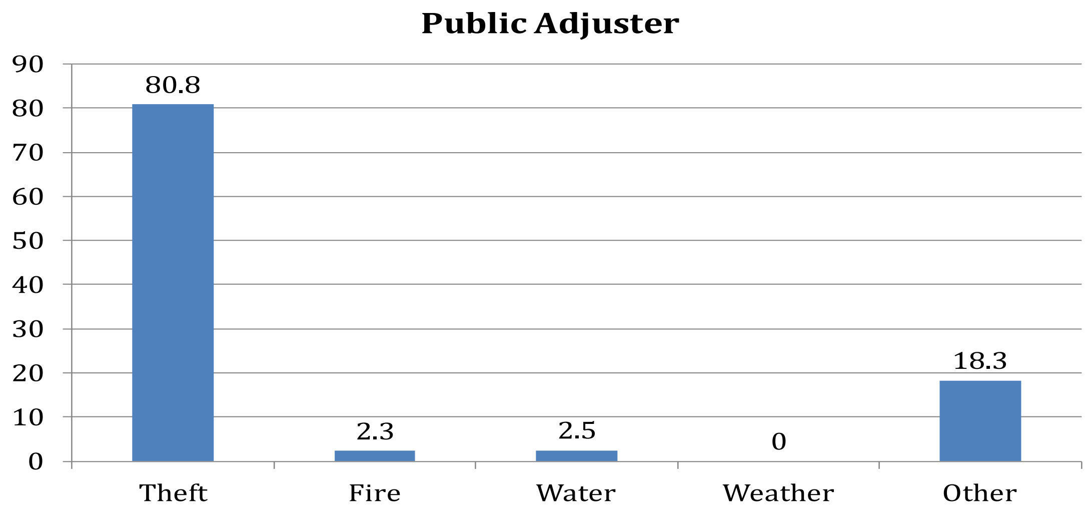
While fraud is a problem that costs insurance companies millions of dollars every year, the majority of claims any particular company handles will be almost or entirely legitimate. For this reason, the claims that are different or the outliers represent the best place to begin to investigate for suspicious activity. As can be seen from the chart above, the outlier clusters (those with the highest distance to nearest cluster – the left side of the x-axis) can be identified and investigated further to determine if they represent suspicious activities. This cumulative distribution can also be used to identify the most suspicious claims and prioritize them for further investigation.

Potentially suspicious clusters can also be identified by looking at the root mean square error of each cluster. The root mean square error is a measure of how variable the claims are within a given cluster. This can identify suspicious claims because the higher the root mean square error, the more dissimilar the claims are, and the fact that a group of more dissimilar claims end up in the same cluster may show evidence that they do not fit well anywhere and therefore might be suspicious. If the claims are different from each other yet wound up in a cluster together, they can be investigated as outliers.

The next approach to identifying suspicious claims within the clustering analysis is to identify claims that are outliers relative to the other claims in its cluster. The cluster as a whole may seem reasonable, but there can be a number of claims within that cluster that are significantly different than the typical claim in that cluster. This will help identify claims that might not seem out of line as a whole, but based on select characteristics, are different than claims with smaller characteristics.

For each claim in a cluster, the distance of the claim from the mean of the cluster is calculated. Claims with a higher distance from the cluster mean are most dissimilar from the average claim in the cluster, yet do not fit better into any other cluster. In order to standardize this measure, the distance from the mean from the cluster analysis can be used as the target in a linear regression model, and a scorecard was developed for a predicted distance from the mean based on the characteristics of the claim. Once a predicted distance is calculated, this distance can be translated into a percentile based on the distribution of the claim distances from means.

The chart below shows the average distance from the cluster mean for claims in a homeowner analysis which were handled by a public adjuster instead of a company adjuster.



As can be seen in the chart, when a public adjuster is involved in a theft claim, the distance from the mean of the cluster is significantly higher for theft claims, and also for the claims for all the other causes of loss combined (Other) . For fire, water, and weather claims, the presence of a public adjuster is not a significant factor.

Those clustering approaches are an automated way to put a list of claims into order for review, allowing a company to prioritize its resources.

Combination of Analytics and Adjuster Experience

Adjuster experience and analytics can be combined using a procedure called PRIDIT, which is the Principal Components Analysis of RIDIT scores. The PRIDIT procedure is described below, and more detail can be found in the referenced paper by Brocket et al[6].

PRIDIT is a mathematical technique using a priori classification of dataset characteristics to score a dataset when there is no definitive training data available. The first step in the PRIDIT analysis is to compile a series of univariate exhibits of suspected predictor variables. Note that without a target, the univariate need only contain an exposure amount for each level, which is claim count. Next, each level of the variable was ranked from 1 to n, with 1 being the most likely characteristic to be fraudulent (this is where the adjuster experience is incorporated). Each level should have a ranking, although multiple levels may be ranked equally. Each level's rank along with the exposure distribution is used to calculate a series of Betas for the variable. A level's Beta is calculated by taking the percentage of total exposure which has a lower (more fraudulent) ranking and subtracting the percentage of total exposure which has a higher (less fraudulent) ranking. An example of the Beta calculation is shown below.

[6] Brocket, Patrick L., Richard A. Derrig, Linda L. Golden, Arnold Levine, and Mark Alpert. "Fraud Classification Using Principal Component Analysis of RIDITs." Journal of Risk and Insurance, 69:3, September, 2002.

Factor Level	Public Adjuster	Claim Count Distribution	PRIDIT Rank (lower = more fraud)	Beta
1	0	0.988	2	0.01
2	1	0.012	1	-0.99

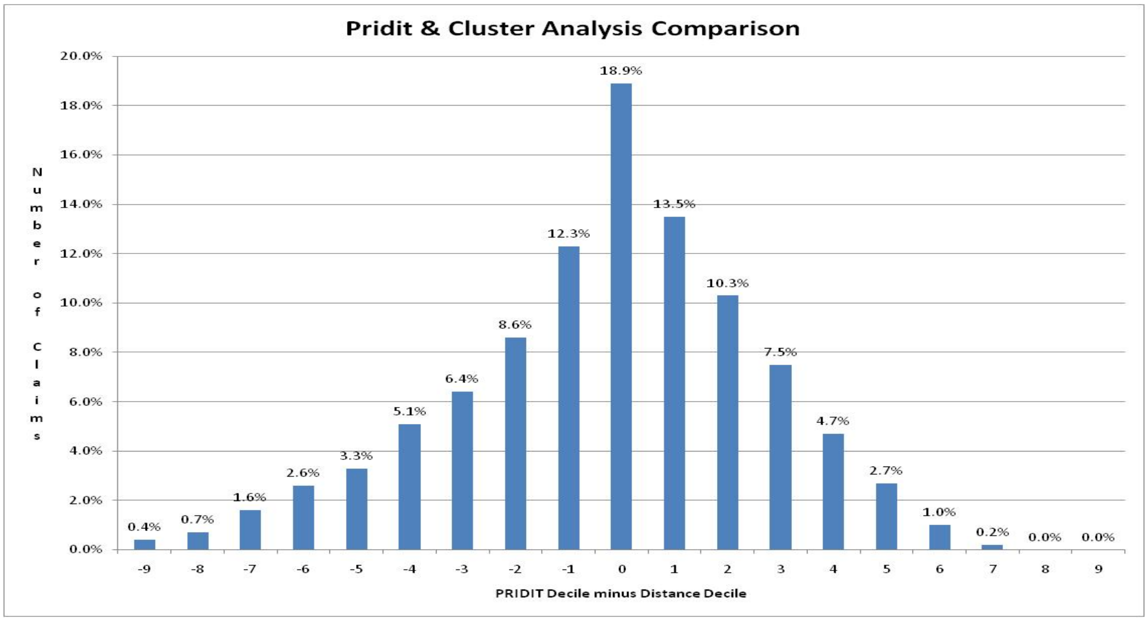
The resulting Betas fall between -1 and 1. This standardization allows variables with varying numbers of levels to be included without skewing the analysis. Betas close to negative one can be interpreted as belonging to a characteristic more likely to be fraudulent as well as a characteristic that is less prevalent in the dataset, which in and of itself is a common red flag in fraud detection.

The individual Betas are then tabulated across variables to produce an overall record Beta. A principal components analysis is run in SAS on the individual claim characteristic Betas in the dataset, and the first principal component is used to standardize these tabulated betas across the dataset and also identify those betas having the largest influence on the fraud calculation. Variables with an absolute value of the weights which are close to one can be interpreted as driving or having the most influence on the PRIDIT fraud score.

Taking these variable weights, we return to the calculation of the overall record Beta in order to obtain the PRIDIT Score. Instead of simply summing the individual variable Betas as was done to obtain the overall record Beta, the PRIDIT Score requires the individual variable Beta to be multiplied by its respective variable weight. The result of this weighting is the record’s PRIDIT Score.

PRIDIT Scores are not predictions. By themselves, a PRIDIT Score can not divulge any information about a record’s fraudulent probability. However, they do provide a way of ranking the claims. Within a dataset, the records with the lowest PRIDIT Scores should be viewed as the most likely to be fraudulent. A company can apply the results of this analysis by directing their limited investigative resources towards claims with the lowest PRIDIT Scores.

Since the PRIDIT analysis is based on suspicion rankings by experienced adjusters, the results can be used as an independent validation of the cluster analysis. To do this, we separated the claims into ten deciles based on predicted suspicion for both the cluster approach and the PRIDIT approach. These deciles were then compared to determine how consistent the two approaches were. The results are shown in the chart below.



As can be seen from the chart, about 64% of the claims are within 2 deciles of each other when comparing the cluster decile to the PRIDIT decile. So a majority of the claims are showing consistent results between the two approaches. However, there are some some claims at either end of the scale that are not consistent. Again, these may be suspicious claims that may not have been identified historically by adjusters, but because of their characteristics are being flagged as potentially suspicious.

APPLICATIONS

Ultimately, the analytics results can be provided to claim adjusters as another tool to assist in identifying suspicious claims. Based on each of the approaches described here, scores can be developed, along with reason codes that identify why a claim is scored at the level it is. An example of this scorecard is shown below. The details of the claim are shown, along with the score and the reason codes.

Claim Details

Arbitration	3	**Accident Date**	10/18/1999
Report Lag	3 days	**Report Date**	10/21/1999
Days Open	932	**Coverage**	Bodily Injury
Lawsuit	Suit Filed		
State	46		
Accident Location	Small Town		
Injury Severity	No Information Available		
Claimant Age	46		

Fraud Model Scores

	Score	Indicator
SIU Referral	53	🟥
Past Identified Fruad	36	🟥
Claim Anomaly	13	🟩
Composite	34	🟨

Fraud Model Reason Codes

1	Delayed Reporting
2	Accident in Small Town
3	
4	

CONCLUSION

Insurance claim fraud is a significant problem, and the focus on claim fraud has increased as well. While insurers have developed effective fraud departments, the reality is that not all fraud is identified by companies, and unidentified fraudulent claims are costing the insurance industry billions of dollars. Predictive analytics can provide increased consistency in the identification of suspicious claims, help companies uncover new types of fraudulent claims, and provide companies with a way to prioritize claims for investigation purposes. By combining transparent statistical techniques with a company's own internal expertise regarding their business, many companies can take significant steps to detecting and preventing fraudulent claims. In today's environment, even a small improvement can have a big impact on a company's bottom line.