**A Prediction System for Detection of Fraud in Auto Insurance Claims Using Machine Learning Techniques**

**Submitted By:**

**Nimesh Katoriwala**

**INSOFE Batch 32, Bangalore**

**PHD HACKATHON**



**International School of Engineering, Bangalore**

**10th February, 2018**

Contents

[**1. Abstract 3**](#_Toc505969899)

[**2. Introduction 3**](#_Toc505969900)

[**3. Why Machine Learning in Fraud Detection ? 4**](#_Toc505969901)

[**4. Exercise Objectives 5**](#_Toc505969902)

[**5. Data Set Description 5**](#_Toc505969903)

[**5.1 Introduction to Datasets 5**](#_Toc505969904)

[**5.2 Challenges Faces in Detection 6**](#_Toc505969905)

[**6. Feature Engineering and Selection 7**](#_Toc505969906)

[**6.1 Feature Engineering 7**](#_Toc505969907)

[**6.2 Importance of Feature Engineering 7**](#_Toc505969908)

[**6.3 Some Engineered Features 7**](#_Toc505969909)

[**7. Model Building and Comparison 8**](#_Toc505969910)

[**7.1 Logistic Regression 9**](#_Toc505969911)

[**7.2 Decision Tree (C50) 9**](#_Toc505969912)

[**7.3 Decision Tree (CART) 10**](#_Toc505969913)

[**7.3 Random Forest 10**](#_Toc505969914)

[**8. Evaluation of Model Performance 11**](#_Toc505969915)

[**8. Model Performance Criterion 11**](#_Toc505969916)

[**9. Model Verdict 12**](#_Toc505969917)

[**9. Key Takeaways 12**](#_Toc505969918)

[**10. Conclusion 13**](#_Toc505969919)

# 1. Abstract

The purpose of this project is to work on Insurance Claims Dataset and to detect the Fraud Claims Insurance fraud has existed since the beginning of insurance as a commercial enterprise. Fraudulent claims account for a significant portion of all claims received by insurers, and cost billions of dollars annually. Types of insurance fraud are diverse, and occur in all areas of insurance. Insurance crimes also range in severity, from slightly exaggerating claims to deliberately causing accidents or damage. Fraudulent activities affect the lives of innocent people, both directly through accidental or intentional injury or damage, and indirectly as these crimes cause insurance premiums to be higher. Insurance fraud poses a significant problem, and governments and other organizations make efforts to deter such activities.

# 2. Introduction

Insurance fraud refers to any claim with the intent to obtain an improper payment from an insurer. Motor and health insurance are the two prominent segments that have seen a spurt in fraud. Frauds can be classified from source and/or nature point of view. Sources can be policyholder, intermediary and/or internal with the latter two being more critical from internal control framework point of view. Frauds can be classified into nature wise, for example, application, inflation, identity, fabrication, staged/contrived/induced accidents etc.

Insurance frauds cover the range of improper activities which an individual may commit in order to achieve a favorable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and ­finally the extent of damage caused.

Potential situations could include:

* Covering-up for a situation that wasn’t covered under insurance¬ (e.g. drunk driving, performing risky acts, illegal activities etc.)
* Misrepresenting the context of the incident: This could include¬ transferring the blame to incidents where the insured party is to blame, failure to take agreed upon safety measures.
* Inflating the impact of the incident: Increasing the estimate of loss¬ incurred either through addition of unrelated losses (faking losses) or attributing increased cost to the losses.

The insurance industry has grappled with the challenge of insurance claim fraud from the very start. On one hand, there is the challenge of impact to customer satisfaction through delayed payouts or prolonged investigation during a period of stress. Additionally, there are costs of investigation and pressure from insurance industry regulators.

On the other hand, improper payouts cause a hit to profitability and encourage similar delinquent behavior from other policy holders. According to FBI, the insurance industry in the USA consists of over 7000 companies that collectively received over $1 trillion annually in premiums. FBI also estimates the total cost of insurance fraud (non-health insurance) to be more than $40 billion annually.

It must be noted that insurance fraud is not a victimless crime – the losses due to frauds, impact all the involved parties through increased premium costs, trust defi­cit during the claims process and impacts to process efficiency and innovation.

Hence the insurance industry has an urgent need to develop capability that can help identify potential frauds with a high degree of accuracy, so that other claims can be cleared rapidly while identifi­ed cases can be scrutinized in detail.

# 3. Why Machine Learning in Fraud Detection ?

The traditional approach for fraud detection is based on developing heuristics around fraud indicators. Based on these heuristics, a decision on fraud would be made in one of two ways. In certain scenarios rules would be framed that would defi­ne if the case needs to be sent for investigation. In other cases, a checklist would be prepared with scores for the various indicators of fraud. An aggregation of these scores along with the value of the claim would determine if the case needs to be sent for investigation. The criteria for determining indicators and the thresholds will be tested statistically and periodically recalibrated.

The challenge with the above approaches is that they rely very heavily on manual intervention which will lead to the following limitations:

* Constrained to operate with a limited set of known parameters based on heuristic knowledge – while being aware that some of the other attributes could also influence decisions.
* Inability to understand context-speci­fic relationships between parameters (geography, customer segment, insurance sales process) that might not reflect the typical picture. Consultations with industry experts indicate that there is no ‘typical model’, and hence challenges to determine the model specifi­c to context.
* Recalibration of model is a manual exercise that has to be conducted periodically to react changing behavior and to ensure that the model adapts to feedback from investigations. The ability to conduct this calibration is challenging.
* Incidence of fraud (as a percentage of the overall claims) is low - typically less than 1% of the claims are classifi­ed. Additionally new modus operandi for fraud needs to be uncovered on a proactive basis.

These are challenging from a traditional statistics perspective. Hence, insurers have started looking at leveraging machine learning capability. The intent is to present a variety of data to the algorithm without judgment around the relevance of the data elements. Based on identi­fied frauds, the intent is for the machine to develop a model that can be tested on these known frauds through a variety of algorithmic techniques.

# 4. Exercise Objectives

Explore various machine learning techniques to improve accuracy of detection in imbalanced samples. The impact of feature engineering, feature selection and parameter tweaking are explored with objective of achieving superior predictive performance.

As a procedure, the data will be split into three different segments – training, validation and testing. The algorithm will be trained on a partial set of data and parameters tweaked on a validation set. This will be examined for performance on the actual testing set. The high-performing models will be then tested for various random splits of data to ensure consistency in results.

In this exercise, I applies a combination of pre-defi­ned rules and predictive machine learning algorithms to identify outliers in data. It is built on Open Source with a library of pre-built algorithms that enable rapid deployment, and can be customized and managed. The approach is comprised of three layers as indicated below:

**OUT COMES**

* Predictions
* Reports
* Case Management

**DETECTION LAYER**

* Business Rules
* ML Algorithms

**DATA HANDLING**

* Data Cleaning
* Transformation
* Sampling

The exercise described above was performed on four different insurance datasets. The Data description for the datasets are given below:

# 5. Data Set Description

## 5.1 Introduction to Datasets

There are total 5 datasets in each Train and Test dataset folders. In Train dataset folders consist of 5 dataset having name as "Train", "Train\_Vehicle", "Train\_Claim", "Train\_Policy", and "Train\_Demographics". Test dataset folders contains 5 sub folders having name as "Test", "Test\_Vehicle", "Test\_Claim", "Test\_Policy", and "Test\_Demographics".

Each Train dataset has a common column having name is "CustomerID" All datasets pertain to claims from a single area and relate to motor/ vehicle insurance claims. In all the datasets, a small proportion of claims are marked as known frauds and others as normal. It is expected that certain claims marked as normal might also be fraudulent, but these suspicions were not followed through for multiple reasons (time delays, late detection, constraints of bandwidth, low value etc.). After merging all Train and Test Datasets, getting below one dataset.

|  |  |
| --- | --- |
|  | **Train data** |
| **No. of records** | 28836 |
| **No. of features** | 42 |
| **Normal Claim** | 21051 |
| **Fraud Claims** | 7785 |
| **Missing Values** | 25776 |

|  |  |
| --- | --- |
|  | **Test data** |
| **No. of records** | 9662 |
| **No. of features** | 41 |
| **Missing Values** | 8668 |

## 5.2 Challenges Faces in Detection

Fraud detection in insurance has always been challenging, primarily because of the skew that data scientists would call class imbalance, i.e. the incidence of frauds is far less than the total number of claims, and also each fraud is unique in its own way. Some heuristics can always be applied to improve the quality of prediction, but due to the ever evolving nature of fraudulent claims intuitive scorecards and checklist- based approaches have performance constraints.

Another challenge encountered in the process of machine learning is handling of missing values and handling categorical values. Missing data arises in almost all serious statistical analyses.

The other challenge is handling categorical attributes. This occurs in the case of statistical models as they can handle only numerical attributes. So, all the categorical attributes are transposed into multiple attributes with a numerical value imputed. For example – the gender variable is transposed into two different columns say male (with value 1 for yes and 0 for no) and female. This is only if the model involves calculation of distances (Euclidean, Mahalanobis or other measures) and not if the model involves trees.

A specific challenge with "Train\_Vehicle" and "Test\_Vehicle" is that it contains duplicate customer id. Each data set contains duplicate customer id having different values. For that Using cast function I reshape the both data frame and convert into column wise.

# 6. Feature Engineering and Selection

## 6.1 Feature Engineering

Success in machine learning algorithms is dependent on how the data is represented. Feature engineering is a process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model performance on unseen data. Domain knowledge is critical in identifying which features might be relevant and the exercise calls for close interaction between a loss claims specialist and a data scientist.

## 6.2 Importance of Feature Engineering

* Better features results in better performance: The features used in the model depict the classi­fication structure and result in better performance.
* Better features reduce the complexity: Even a bad model with better features has a tendency to perform well on the datasets because the features expose the structure of the data classi­fication.
* Better features and better models yield high performance: If all the features engineered are used in a model that performs reasonably well, then there is a greater chance of highly valuable outcome.

## 6.3 Some Engineered Features

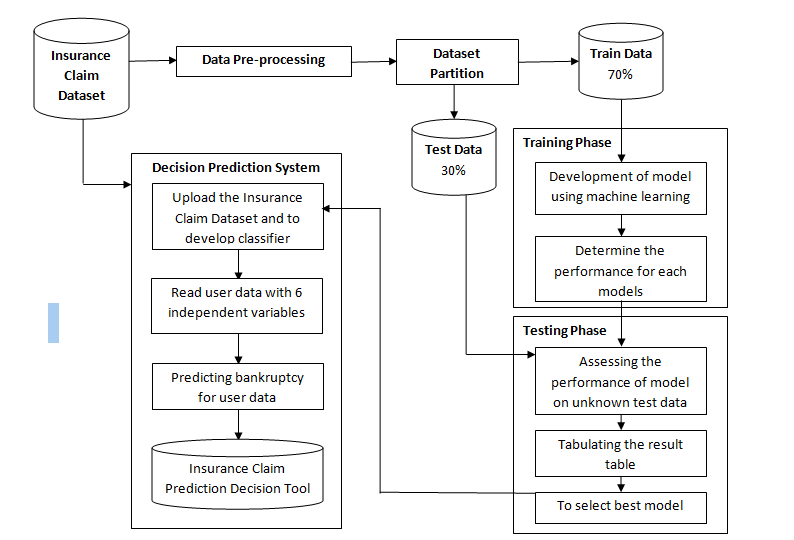
Lot of features are engineered based on domain knowledge and dataset attributes, some of them are listed below:

* Remove those column which are not redundant, also remove those column which gives less information as well less frequent columns.
* Finding a new column "TotalnoofDays" by subtracting two dates.
* Find out the "VehicleAge" from dataset.
* Splitting the "Policy\_CombinedSinglelimit" into "Policy\_SplitLimit" and "Policy\_CombinedsingleLimit".
* Identifying "FinancialStatus" from "CapitalGain" and "CapitalLoss" column.

# 7. Model Building and Comparison

The model building activity involves construction of machine learning algorithms that can learn from a dataset and make predictions on unseen data. Such algorithms operate by building a model from historical data in order to make predictions or decisions on the new unseen data. Once the models are built, they are tested on various datasets and the results that are considered as performance criterion are calculated and compared.

The Methodology of this project is illustrated in below figure:

****

The following steps are listed to summarize the process of the model development:

* Once the dataset is obtained and cleaned, different models are tested on it.
* Based on the initial model performance, different features are engineered and re-tested
* Once all the features are engineered, the model is built and run using different β values and using different iteration procedures (feature selection process)
* In order to improve model performance, the parameters that affect the performance are tweaked and re-tested

Separate models generated for each fraud type which self-calibrate over time - using feedback, so that they adapt to new data and user behavior changes. Multiple models are built and tested on the above datasets. Some of them are:

## 7.1 Logistic Regression

It is a classifier that predicts the outcome based probabilities of logistic function. It estimates the relationship between different independent variables and the dependent variable based on probabilistic value. It may be either binary or multinomial classifier.

The logistic function is denoted as:

**C:\Users\NY 5211\Desktop\logistic.PNG**

β0 is an intercept and β's are coefficients for input variables X. The value of F(X) ranges from zero to one. The logistic regression model generated is also called as generalized linear model.

**Advantages:**

* Low variance
* Provides probabilities for outcomes

**Disadvantages:**

* Cannot handles non-linearity in the data
* Restrictions on the dependent variable, can take discrete values only
* Requires large sample size to predict/perform better.

## 7.2 Decision Tree (C50)

C5.0 algorithm is widely used as a decision tree method in machine learning. Initially we have ID3.0 algorithm. Based on ID3.0, people developed C4.5 algorithm, and finally develop C5.0 algorithm. This type of decision tree model is based on entropy and information gain.

The decision tree based on entropy and information gain (ID3, C4.5, C5.0) is to build a tree by partitioning with highest information gain. For categorical variables, we partition by each level and find the best variable with highest information gain. For continuous variables, we find the best cut point to do a binary partition with highest information gain. For missing data, we create a separate node, which is not causing changes of information gain.

**Advantages:**

* Simple, easy to understand
* Less Data cleaning
* Data type is not constraint

**Disadvantages:**

* Over fitting
* Not fit for Continuous variable

## 7.3 Decision Tree (CART)

Classification and Regression tree or CART is a type of supervised learning algorithm which is works for both classification and regression. In this technique we split the population/sample based on most significant input variables. In CART node splitting based on GINI index and Information Gain. which formulate as follows:

GINI INDEX:

G = 1 - sum(pi) 2

Information Gain:

I = Gbefore - Gafter

**Advantages:**

* Simple, easy to understand
* Less Data cleaning
* Data type is not constraint

**Disadvantages:**

* Over fitting
* Not fit for Continuous variable

## 7.3 Random Forest

Random forest are classifier which construct decision trees for building the model and outputs the mode value of individual trees as result of prediction. Classification is performed by selecting a new input vector from training set. The vector is placed at the bottom of each of the trees in the forest. The proximity is computed for the tree. If the tree branches are at the same level, then proximity is incremented by one. The proximity evaluated is standardized as a function of the number of trees generated. Random forest compute the important features in a dataset based on the out of bag error estimate. The algorithm also reduces the rate of over fitting observed in decision tree models.

**Advantages:**

* De-correlates the models in the ensemble.
* Improve accuracy of prediction

**Disadvantages:**

* It prevents Overfitting
* Unlike Decision trees, they don't give you a lot of insight

# 8. Evaluation of Model Performance

The algorithm were initially applied for the training set to develop predictive models. These models were further evaluated using the test set. Performance of each model was adjudged using different statistical parameters like confusion matrix and Receiver Operating Characteristics (ROC) curve. Confusion matrix is a contingency table that represents the performance of machine learning algorithms. It represents the relationship between actual class outcome and predicted class outcome based on the following four estimates:

1. **True Positive**: The actual negative class outcome is predicted as negative class from the model
2. **False Positive**: The actual negative class outcome is predicted as a positive class outcome. It leads to TYPE - 1 error.
3. **False Negative**: The actual positive class outcome is predicted as negative class from the model. It leads to TYPE - 2 error.
4. **True Negative**: The actual class outcome excluded is also predicted to be excluded from the model.

Based on these four parameters the performance of algorithm can be judged by calculating the following ratios.

1. **Accuracy:**
2. **Recall:**
3. **Precision:**
4. **F1 Score:**

2 \*

# 8. Model Performance Criterion

The Model performance can be evaluated using different measures. Some of them used in this project are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Recall(%)** | **Precision(%)** | **F1 Score(%)** |
| Random Forest | 79 | 90 | 83 |
| GBM | 74 | 89 | 81 |
| Decision Tree(C50) | 66 | 79 | 71.86 |
| Decision Tree(CART) | 65 | 71 | 68 |
| Logistic Regression | 45 | 77 | 56 |

# 9. Model Verdict

In this analysis, many factors were identifi­ed which allows for an accurate distinction between fraudulent and honest claimants i.e. predict the existence of fraud in the given claims. The machine learning models performed at varying performance levels for the different input datasets. By considering the average F5 score, model rankings are obtained- i.e. a higher average F5 score is indicative of a better performing model.

The analysis indicates that the modi­fied random under Sampling and Adjusted Random Forest algorithms perform best. However, it cannot be assumed that the order of predictive quality would be replicated for other datasets. As observed in the dataset samples, for feature rich datasets, most models perform well Likewise, in cases with signifi­cant quality issues and limited feature richness the model performance gets degraded.

# 9. Key Takeaways

* **Predictive quality depends more on data than on algorithm:**

Many researches indicate quality and quantity of available data has a greater impact on predictive accuracy than quality of algorithm. In the analysis, given data quality issues most algorithms performed poorly on dataset. In the better datasets, the range of performance was relatively better

* **Poor performance from logistic regression**

Logistic regression is more of a statistical model rather than a machine earning model. It fails to handle the dataset if it is highly skewed. This is a challenge in predicting insurance frauds, as the datasets will typically be highly skewed given that incidents will be low. MMVG is built with an assumption that the input data supplied is of Gaussian distribution which might not be the case. It also fails to handle categorical variables which in turn are converted to binary equivalents which leads to creation of dependent variables.

* **Outperformance by Ensemble Classi­fiers:**

Both the boosting and bagging being ensemble techniques, instead of learning on a single classi­er, several are trained and their predictions are aggregated. Research indicates that an aggregation of weak classifi­ers can out-perform predictions from a single strong performer.

* **Loss of Intuition with Ensemble Techniques**

A key challenge is the loss of interpretability because the ­nal combined classi­er is not a single tree (but a weighed collection of multiple trees), and so people generally lose visual clarity of a single classi­fication tree. However, this issue is common with other classifi­ers like SVM (support vector machines) and NN (neural networks) where the model complexity inhibits intuition. A relatively minor issue is that while working on large datasets, there is signi­ficant computational complexity while building the classi­er given the iterative approach with regard to feature selection and parameter tweaking. Anyhow given model development is not a frequent activity this issue will not be a major concern.

# 10. Conclusion

The machine learning models that are discussed and applied on the datasets were able to identify most of the fraudulent cases with a low false positive rate i.e. with a reasonable precision. This enables loss control units to focus on new fraud scenarios and ensuring that the models are adapting to identify them. Certain datasets had severe challenges around data quality, resulting in relatively poor levels of prediction.

Given inherent characteristics of various datasets, it would be impractical to a' priori de­fine optimal algorithmic techniques or recommended feature engineering for best performance. However, it would be reasonable to suggest that based on the model performance on back-testing and ability to identify new frauds, the set of models offer a reasonable suite to apply in the area of insurance claims fraud. The models would then be tailored for the speci­fic business context and user priorities.