Insurance Claim Fraud Detection Report

Submitted By:

Sameer Kumar Ramakrishnagari

Contents

Insurance Claim

INTRODUCTION

For years now, we've been told that data is king and that it should be tapped for all decisions: what to stock, how much to buy, what products to suggest repeating customers. There are numerous ways data has been employed in retail. All retailers want to know their target buyer, but understanding the past and present of their interactions simply isn't enough. The next piece of the puzzle is being able to project what customers will do and need next. Machine learning in retail pieces together the fragmented puzzle we've been looking at for years. It accomplishes this by combining customer data with market trends to give retailers a holistic action plan to target customers better and ultimately achieve their goal to drive revenue growth in a more efficient way.

This study focuses on Insurance Claim. An **Insurance Claim** is a formal request to an insurance company for the coverage or compensation for a covered loss or policy event. The insurance company validates the claim and once approved, issues payment to the insured or an approved interested party on behalf of the insured. In many cases, the third parties file claims on behalf of the insured person but generally only the person listed in the policy is entitled to claim payments.

I. a. Business Case Statement

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.
Fraud affects the lives of innocent people as well as the insurance industry and thus it may be of interest for the health of the Insurance Industry and Society. In fact, Insurers report certain classified cases to Regulator and Law enforcement agencies like Police, Crime Bureaus and others as mandated by the Regulators/Government and required by Law. With the advent of organised gangs and/or collusion, the problem has become more complex and sophisticated and the frauds have been difficult to detect and to prove, if detected.

I. b. Problem Statement

At the Regulator and Law enforcement level, the intelligence arising out of prediction will help revamp the Regulations/Laws and plan not only enforcement but Industry based initiatives/systems for resilience and to share information for consumption of the Industry and the Society.
Prediction at the time of processing claims will reduce costs and minimize losses for the insurance company. Hence, prediction of fraud plays very important role in auto insurance claims. The

company wants to understand the hidden patterns in the data which lead to construction of investigation process as well as claim settlement decision.

II. a. Data Understanding

- **Claim Report Data**

This data set representing the events that has occurred as a part of insurance claim. It includes Date & Time, Severity, Location of incident, Intermediaries involved in the incident and the monetary data representing the Claim amount.

- **Customer Demographic Data:**

This data set provides the details of the Customer who has opted for an insurance Claim. The data set includes, the customer's age, gender, location, financial status, education level and occupation of the 1st party insurer.

- **Policy Data:**

This data set provides a brief idea of the auto insurance policy taken by the customer. It includes, Where and When the policy was taken, premium to be paid, Umbrella Amount – Representing any upper bound insurance limit that the insurance company will pay in the event of the Customer unable to repay the loss amount.

- **Vehicle Data:**

This one is a tricky data set. The record set count is a quadruple of other data set. The reason being, the vehicle attributes are in row format instead of the general columnar format. Also, it includes, which vehicle was involved in the accident, type and manufacturer of the vehicle

- **Classified Transaction Data**

This is a dataset indicating whether a customer transaction is Authentic or Fraudulent.

II. b. Handling Missing data

- **Claim Data:**

The Claim data needs a wide range of analysis on the missing data. The Amount of Total Claim column has 50 Null values. However, taking a closer look on the data shows that the sum of Amount of Injury Claim, Amount of Property Claim and the Amount of Vehicle Damage is exactly equal to the Amount of Total Claim. Hence the missing values need to be imputed based on the sum of other three columns. Moving further, the Property Damage and the Police Report has majority of the missing values in the data. It is quite difficult to bucket these values by means of imputation. Further, since there this accounts to more than 35% of the missing values, rather than imputing the column, eliminating the column makes sense. The Type of Collision column has almost 5000 missing values and it is spread across 3 different categories. Hence rather than imputing it with the MODE value, created a separate category for this column. Finally, the incident time and witnesses has been imputed with Median and Mode respectively.

- **Customer Demographic Data:**

The customer data has NA values in the Insured Gender and the Country Columns. While looking at the distribution of the Gender, it is spread across both the genders with a slightly higher number of Female gender. Hence, impute the Gender with Female for the 30 missing records. While looking at the Country, it is clear that there are 2 missing values and other data belongs to a single Country India. Thus the other two missing records has been imputed with Country = India.

- **Policy Data:**

The policy data has missing values in the policy premium column. However, this being a numeric data, imputing the missing values with MODE will not work. There are multiple ways to impute this data viz. Mean, Median, KNN Imputation, and Central Imputation. Upon checking the box plot, there are multiple outliers that exists in the Policy Premium. However, while looking at the business case, policy premium is an amount paid for the policy taken. Hence, ignoring the outliers, central imputation has been done.

- **Vehicle Data:**

As stated in the data understanding step, Vehicle data set contains redundant Customer data as the data is spread across a row format and after converting it into a columnar format, it can be observed that the Vehicle Make Column contains Null Values. This data is similar to the Total Claim Amount. The Vehicle Make Column needs imputation based on the Vehicle Model and the same has been done in the data set.

- **Claim Report Data**

This data has quite a good number of missing values. Property Damage column has more than 10

II. c. Outlier Detection

A *box plot* is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum. The top of the rectangle indicates the third quartile, a horizontal line near the middle of the rectangle indicates the median, and the bottom of the rectangle indicates the first quartile. A vertical line extends from the top of the rectangle to indicate the maximum value, and another vertical line extends from the bottom of the rectangle to indicate the minimum value.

Majority of the data set contains categorical columns and only a very few Numerical Columns exists of most of which are Amount pertaining to the insurance Claim. A box plot on the Numerical Column provides better insight of the data. Outliers do exists in the Umbrella Amount, Total Claim, and Vehicle Damage and Property Damage columns. However, looking at the business case, these values are specific to the vehicle type and the insurance company and hence no processing has been done on the Outliers

II. d. Correlation

All correlations have two properties: **strength** and direction.

- The strength of a correlation is determined by its numerical value.
- The direction of the correlation is determined by whether the correlation is positive or negative.

- Positive correlation: Both variables move in the same direction. As one variable decreases, the other variable also decreases.
- Negative correlation: The variables move in opposite directions. As one variable increases, the other variable decreases.

A general intuition is that, a correlation must exist on the Amount of Total Claim, Amount of Property Damage, Amount of Injury and Amount of Vehicle Damage as the former is a total sum of the latter 3 attributes. A corrplot on these three columns depicts the same. Since Amount of Total Claim is dependent on the other 3 Amount, as a process of feature reduction, the other three columns has be removed too.

Similarly, correlation exists between the age of the person and the customer loyalty period and hence, retaining either of the column should suffice and in the journey of modelling, the Customer Loyalty Period has been taken into account leaving the other.

II. e. Feature Creation

Dates form a major role in the feature creation and feature reduction process in the data set. The incident date and the policy coverage date are the two columns which can be split into years, months and days. Also, it can be inferred that the incident records the same year and hence incident year can be removed from the data set. Hence dates provide both feature creation in the data set.

III. Exploratory Data Analysis - Classification Problem

a. Logistic Regression

The name ‘Regression’ here implies that a linear model is fit into the feature space. This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables. The odds or probabilities that describe the outcome of a single trial are modelled as a function of explanatory variables. Logistic regression algorithms helps estimate the probability of falling into a specific level of the categorical dependent variable based on the given predictor variables.

In our business problem we want to predict if our customer will churn or not. Here the outcome of the prediction is not a continuous number because there will either be attrition or no attrition and hence linear regression cannot be applied. Here the outcome variable is one of the several categories and using logistic regression helps.

Binary Logistic Regression – The most commonly used logistic regression when the categorical response has 2 possible outcomes i.e. either yes or not. This approach is suitable for our problem statement

When to Use Logistic Regression Machine Learning Algorithm: Use logistic regression algorithms when there is a requirement to model the probabilities of the response variable as a function of some other explanatory variable. For example, probability of a customer churning given as a function of revenue generated per customer, frequency of purchase.
Use logistic regression algorithms when there is a need to predict probabilities that categorical

dependent variable will fall into two categories of the binary response as a function of some explanatory variables.

Logistic regression algorithms are also best suited when the need is to classify elements two categories based on the explanatory variable.
# Your answer passed the tests! Your score is 62.36%
# Auxiliary metrics => Precision=60.74587% and Recall=64.07137%

b. Decision Tress (C5.0)

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. C5.0 algorithm to build either a decision tree or a rule set. This type of decision tree model is based on entropy and information gain.

A C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the sub samples cannot be split any further. Finally, the lowest-level splits are re-examined, and those that do not contribute significantly to the value of the model are removed or pruned. The C5.0 node can predict only a categorical target. And we have categorical target to classify thus we go with this and see how it is behaving. In R, the C50 package is used to get C5.0 algorithm in classification.

In contrast, a rule set is a set of rules that tries to make predictions for individual records. Rule sets are derived from decision trees and, in a way, represent a simplified or distilled version of the information found in the decision tree. Rule sets can often retain most of the important information from a full decision tree but with a less complex model. Because of the way rule sets work, they do not have the same properties as decision trees. The most important difference is that with a rule set, more than one rule may apply for any particular record, or no rules at all may apply. If multiple rules apply, each rule gets a weighted "vote" based on the confidence associated with that rule, and the final prediction is decided by combining the weighted votes of all of the rules that apply to the record in question. If no rule applies, a default prediction is assigned to the record.
# Your answer passed the tests! Your score is 89.1%
# Auxiliary metrics => Precision=95.24689% and Recall=83.6983%

c. Decision Tress (rPart)

The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.
Some techniques, often called ensemble methods, construct more than one decision tree:
Boosted trees Incrementally building an ensemble by training each new instance to emphasize the training instances previously mis-modeled. A typical example is AdaBoost. These can be

used for regression-type and classification-type problems.
Bootstrap aggregated (or bagged) decision trees, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction.
Decision trees are formed by a collection of rules based on variables in the modeling data set:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

# Your answer passed the tests! Your score is 89.1%
# Auxiliary metrics => Precision=95.24689% and Recall=83.6983%

d. Random Forest Model:

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART model with different sample and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

Random Forest is the go to machine learning algorithm that uses a bagging approach to create a bunch of decision trees with random subset of the data. A model is trained several times on random sample of the dataset to achieve good prediction performance from the random forest algorithm. In this ensemble learning method, the output of all the decision trees in the random forest, is combined to make the final prediction. The final prediction of the random forest algorithm is derived by polling the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.

When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows: Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.
# Your answer passed the tests! Your score is 48.09%

# Auxiliary metrics => Precision=89.701% and Recall=32.84672%

e. GBM Model

Gradient Boosting Model build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive and the interaction with the higher terms increases. There are typically three parameters - number of trees, depth of trees and learning rate, and each tree built is generally shallow.

# Your answer passed the tests! Your score is 85.74%
# Auxiliary metrics => Precision=87.93691% and Recall=83.65775%

f. Naive Bayes Model

Naive Bayes (NB) is a very simple algorithm based around conditional probability and counting. Essentially, model is actually a probability table that gets updated through the training data. To predict a new observation, you'd simply "look up" the class probabilities in the “probability table" based on its feature values.

It's called "naive" because its core assumption of conditional independence (i.e. all input features are independent from one another) rarely holds true in the real world.

- **Strengths**: Even though the conditional independence assumption rarely holds true, NB models actually perform surprisingly well in practice, especially for how simple they are. They are easy to implement and can scale with your dataset.
- **Weaknesses**: Due to their sheer simplicity, NB models are often beaten by models properly trained and tuned using the previous algorithms listed.

# Your answer passed the tests! Your score is 59.45%
# Auxiliary metrics => Precision=55.21558% and Recall=64.39578%

Conclusion:

C5 is a classifier which classifies the data in less time compare to other classifier. For generating decision tree the memory usage is minimum and it also improve the accuracy. This proposed system is developed on the bases of C5 algorithm. In the proposed system C5.0 algorithm provides Feature selection, Cross validation and reduced error pruning facilities. So the further scope of this algorithm is achieved by implementation of new features like PCA, Cross Validation and Model Complexity.

C5.0 performs well with F1 score of 89.1%.
# Auxiliary metrics => Precision=95.24689% and Recall=83.6983%