# Capstone Project: Improving Insurance Claims Management

Lisa Ang

## Data Story: Introduction

While the concept of shared risk management has a long and notable history in many cultures around the world, the specialized varieties of modern insurance practiced today developed in Europe during the 17th and 18th century with the Age of Enlightenment. Managed prudently, insurance products can play very important roles in society at all levels, from individuals and corporate entities to government. It can enable economic development through the underwriting of trade, while also serving to provide peace of mind to average citizens through collective protection against hazards such as flooding, fire and medical emergencies, or the risks of performing routine activities like driving a vehicle or travel.

From the standpoint of an insurer, claims management is one of the most critical aspects of a functioning insurance provider. To be successful in today’s highly competitive economic environment, the insurer must develop a consistent operating model that can balance claim costs with optimal risk management and client satisfaction, while eliminating unnecessary expenses associated with claims handling. A large part of this involves maximizing the efficiency of the claims process and reducing the need for manual evaluation of claims. Although the execution of claims handling is necessarily highly customized from industry to industry, most elements are quite similar when broken down into their core processes. When combined with a well-designed case management platform, this makes many aspects of claims handling excellent candidates for automation.

For many types of insurance, some claims can be approved with minimal involvement quite early in the claims process but others will require additional information to be obtained prior to approval. Identifying the claims that can be approved quickly is an effective means of streamlining the claims process. This not only reduces costs but leads to greater customer satisfaction and is therefore of great importance to the insurer.

BNP Paribas Cardif is an international insurance company specializing in personal insurance coverage with over 90 million clients in 36 countries across Europe, Asia and Latin America. The insurance claims they receive can vary widely in complexity as well as the levels of verification required before a claim can be approved and payment issued. BNP Paribas Cardif has provided an anonymized database with two types of claims containing data available upon receipt of a claim:

1. Claims which meet requirements for accelerated approval and faster payments
2. Claims for which additional information is required before approval can be given and payment issued

The aim is to determine which claims can be selected for accelerated approval. The database is provided in the form of two anonymized datasets (train.csv and test.csv) for a competition on [Kaggle](https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/submit). The goal for the analysis is to use the training set to generate a model that will predict a probability for each claim in the test set for accelerated approval. The accuracy of the prediction will be evaluated by submission on the Kaggle website and scored by log loss ranking.

## Initial Impressions

### Structure

Upon reading in the data, the training set (train.csv) gives a dataframe of 114321 observations of 133 variables (131 independent variables, target and ID). The test set (test.csv) gives a dataframe of 114393 observations of 132 variables (131 independent variables and ID). Variables are either categorical or numeric. All string type variables are categorical and there are no ordinal variables. Of the 131 independent variables, 19 are categorical, and 112 are numeric.

The dependent variable is the “target” column in the training set, where 1 represents claims suitable for accelerated approval and 0 represents claims requiring additional verification.

### Target imbalance

By analyzing the counts of 1 and 0 target column of the training set, there are 87021 instances of 1 and 27300 instances of 0, indicating that approximately 3 in 4 claims can be flagged for accelerated approval. This results in a modest class imbalance ratio of 3:1, which may bias any models built on the training set towards a result of 1 for accelerated approval, as most conventional machine learning algorithms are biased towards the majority class. There are numerous approaches that can be used to address this issue, such as rebalancing the training data by oversampling the minority class, undersampling the majority class or creating new minority classes from the majority class.

### Column names

Besides target and ID, the independent variables from raw data do not have descriptive column names and are named generically (eg. v1, v2, v3 etc). As it is not possible to impute these, this could present difficulties in employing domain knowledge for feature selection and engineering, as well as determining appropriate methodology for data wrangling.

## Data Preparation

### Data partitioning

As a test set without the dependent target variable has been provided for model evaluation, the training set was partitioned into training and validation sets in a 50:50 split, maintaining the same ratio of the dependent variable in each set. The training set has 57160 observations of 133 variables (131 independent variables, target and ID) while the validation set has 57161 observations of 133 variables (131 independent variables, target and ID).

### Missing values

Upon first inspection of structure and summary, there appear to be many missing values. When these are counted, there are 100 columns with greater than 25% missing values. There are visually obvious trends in many rows with similar missing observations, which may indicate that patterns of missing values could have predictor capabilities. While imputing such a large amount of missing values can sometimes improve results, because descriptive column names are not provided, it is not possible to determine relevance to analysis based on domain knowledge. Thus for purposes of this project, columns containing greater than 25% missing values were removed from the data set.

### Separation of categorical and numeric variables

Data was split into numeric and categorical variables for separate processing of outliers and remaining missing values.

### Categorical variables: Levels and missing values

The number of levels for all categorical variables in training data was determined and all categorical variables with more than 15 levels were removed. This gives 13 remaining categorical variables: v3, v2, v30, v31, v47, v52, v66, v71, v74, v75, v91, v107, v110. The same variables were retained from the validation and test data.

For these remaining variables: In columns with less than or equal to 5% missing values, NA was replaced with the most common level. In columns with greater than 5% (but less than 25%) missing values, NA was replaced with a new level called “missing.” The same was performed for the validation and test sets.

### Numeric variables: Outliers and missing values

For numeric variables, outliers were defined as lesser than the 5th or greater than the 95th percentiles for each column. Outliers were replaced with the value of original mean of the column, ie. the mean of the non-outliers, excluding missing values. For the remaining missing values, NA was replaced with the median of each column.

## Data Exploration

### Variable screening by Information Value

The use of information value (IV) and weight of evidence (WOE) is a very popular screening method for selecting predictor variables for binary classifier models and has been used in the credit scoring world for several decades to classify customer loan repayment probabilities.

IV and WOE facilitate the consideration of each variable’s independent contribution to the outcome. It can detect linear and nonlinear relationships, enables visualization of the correlation between predictive variables and the dependent variable, and provides a means of ranking variables in terms of univariate strength. For the purposes of this project, IV also enables the comparison between continuous and categorical variables without the need to create dummy variables.

Briefly, WOE is used to transform a continuous independent variable into a set of groups based on similarity of dependent variable distribution.6 The continuous independent variables are split into 10 or less bins and WOE calculated for each bin. Adjacent bins with similar scores may be combined. IV can then be calculated for each independent variable.

By convention, the values of the IV statistic can be interpreted by the following table:

Table 1: IV and variable predictiveness

|  |  |
| --- | --- |
| Information Value | Variable Predictiveness |
| < 0.02 | Not useful for prediction |
| 0.02 to 0.1 | Weak predictive power |
| 0.1 to 0.3 | Medium predictive power |
| 0.3 to 0.5 | Strong predictive power |
| > 0.5 | Suspicious predictive power |

Using the information package, IV was calculated for the numeric variables alone, as well as on a combined set of numeric and categorical variables:

Table 2: Numeric variables only

|  |  |
| --- | --- |
| Variable | IV |
| v50 | 0.46078426 |
| v129 | 0.17578331 |
| v10 | 0.16512801 |
| v62 | 0.15019944 |
| v14 | 0.13765286 |
| v12 | 0.08555659 |
| v21 | 0.08513701 |
| v34 | 0.05674745 |
| v38 | 0.05434840 |
| v114 | 0.03370896 |
| v72 | 0.02830984 |
| v40 | 0.01364609 |

Table 3: Top 15 Numeric and categorical variables

|  |  |
| --- | --- |
| Variable | IV |
| v50 | 0.467982813 |
| v31 | 0.226882245 |
| v129 | 0.177234292 |
| v47 | 0.160790125 |
| v62 | 0.157678475 |
| v110 | 0.152982998 |
| v66 | 0.129938822 |
| v12 | 0.09030852 |
| v21 | 0.088635846 |
| v38 | 0.049081143 |
| v114 | 0.032781355 |
| v72 | 0.031632212 |
| v30 | 0.013698502 |
| v24 | 0.011741035 |
| v71 | 0.007115383 |

Based on these rankings of each variable’s independent contribution to the outcome of the target variable, the original data set has been reduced to a manageable number of features with significant predictor capabilities. It should be possible to proceed with the original approach and build either logistic regression and/or decision tree models based on variables with medium and strong predictive power.