# Capstone Project: Improving Insurance Claims Management

Lisa Ang

## Data Wrangling

This is a summary of the data preparation performed on the BNP Paribas Cardif dataset prior to exploratory analysis.

#### 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. This reduced the training set to 57160 observations of 32 variables including target. The same columns were removed from the validation and test sets.

#### 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.