# Capstone Project: Improving Insurance Claims Management

Lisa Ang

## Using machine learning to identify claims for accelerated approval

The BNP Paribas Cardif database is an anonymized dataset with two types of claims containing data available at an early stage of the claims handling process:

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

As the goal is to determine which claims can be selected for accelerated approval based on patterns detected in training data for which the dependent variable “target” is supplied, this is a supervised classification problem, where the relationship between variables will be used to generate an algorithm that can learn from the training data and make predictions for out-of-sample data.

Logistic regression is a statistical method that is commonly used for analyzing data in which there are one or more independent variables that determine a dichotomous characteristic or outcome. The goal of logistic regression is to find the best-fitting model to describe the relationship between a binary response variable and the predictor variables. Thus, it is well suited for the purpose of this project.

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 will be used for variable ranking. WOE may be used for imputation. By convention, only variables with medium or strong predictive power (IV 0.1-0.5) will be used to build the models.

|  |  |
| --- | --- |
| Variable | IV |
| v50 | 0.467982813 |
| v31 | 0.226882245 |
| v129 | 0.177234292 |
| v47 | 0.160790125 |
| v62 | 0.157678475 |
| v110 | 0.152982998 |
| v66 | 0.129938822 |

When developing models to be used for prediction, the most critical metric is necessarily how well the model performs in predicting the target variable on out-of-sample data. The training dataset provided by BNP Paribas was previously split into training and validation sets for this purpose. To determine accuracy, the models will be first evaluated by Area Under Curve (AUC) on the validation set prior to log loss ranking of test set predictions on the Kaggle website.