# Capstone Project: Improving Insurance Claims Management

Lisa Ang

## Statistics and Data Exploration

### Histograms

To visually explore the remaining numeric and categorical variables for any obvious patterns of variation, histograms of each variable was plotted with ggplot, faceted by the dependent target variable. Based on visual observation, there are variables that may have zero or near zero variance.

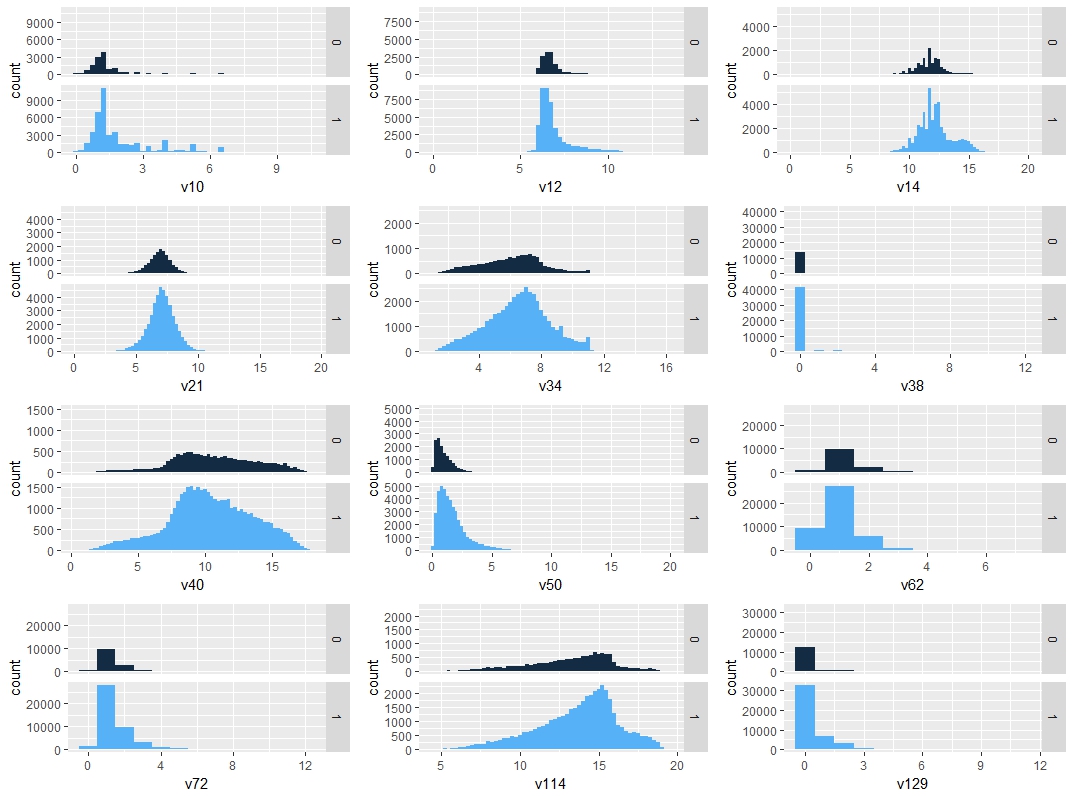
****

Figure 1: Distribution of numeric variables v10, v12, v14, v21, v34, v38, v40, v50, v62, v72, v114, v129

****

Figure 2: Distribution of categorical variables v3, v24, v30, v31, v47, v52, v66, v71, v74, v75, v91, v107, v110

### Correlation analysis of numeric variables

To estimate multicollinearity, a correlation matrix was generated for the remaining 12 numeric variables (v10, v12, v14, v21, v34, v38, v40, v50, v62, v72, v114, v129) to determine if any variables demonstrated a linear dependence that could contribute to overfitting during modelling. Using a pairwise absolute correlation cutoff of 0.75, 4 variables (v10, v14, v34, v40) were identified.

### Zero and near-zero variance

Variables with a single unique value across the data set are typically uninformative and may present problems with model fitting. These can usually be safely removed; however, predictors with very small, non-zero variance across the samples are not necessarily uninformative and may contribute greatly to the model’s accuracy. The remaining 12 numeric and 13 categorical variables were analyzed for zero and near-zero variance using the nearZeroVar() from the caret package. No variables had zero variance, but three were identified with variance less than 10%:

Table : Near-zero variance predictors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | freqRatio | percentUnique | zeroVar | nzv |
| v38 | 52.42693 | 0.020994 | FALSE | TRUE |
| v3 | 504.6549 | 0.005248 | FALSE | TRUE |
| v74 | 153 | 0.005248 | FALSE | TRUE |

### 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.

As IV and WOE were used almost exclusively for credit scoring purposes in its development, the terms “good” and “bad” were historically used for non-events and events, hence IV is defined as:

IV = \sum (Distribution Good_{i}-Distribution Bad_{i})\times WOE_{i} 

where:

Weight of Evidence = ln(\frac{Distribution Good_{i}}{Distribution Bad_{i}})   

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 : 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, both with and without correlated numeric variables identified during correlation analysis. Results were as follows, with highly correlated variables shown in yellow:

Table : Numeric variables only, without correlated variables

|  |  |
| --- | --- |
| Variable | IV |
| v50 | 0.46078426 |
| v129 | 0.17578331 |
| v62 | 0.15019944 |
| v12 | 0.08555659 |
| v21 | 0.08513701 |
| v38 | 0.05434840 |
| v114 | 0.03370896 |
| v72 | 0.02830984 |

Table : Numeric variables only, with correlated variables

|  |  |
| --- | --- |
| 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 : Top 15 Numeric and categorical variables, without correlated variables

|  |  |
| --- | --- |
| Variable | IV |
| v50 | 0.46078430 |
| v31 | 0.23007160 |
| v129 | 0.17578330 |
| v47 | 0.15995130 |
| v110 | 0.15103240 |
| v62 | 0.15019940 |
| v66 | 0.12845110 |
| v12 | 0.08555659 |
| v21 | 0.08513701 |
| v38 | 0.05434840 |
| v114 | 0.03370896 |
| v72 | 0.02830984 |
| v24 | 0.01337383 |
| v30 | 0.01130699 |
| v74 | 0.00765538 |

Table : Top 15 Numeric and categorical variables, with correlated 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 |