**Capstone Project: Improving Insurance Claims Management**

Accelerating claims approval for BNP Paribas Cardif

4 November 2017

Lisa S. Ang

Table of Contents

[1. Introduction 3](#_Toc497334125)

[2. Obtaining data: BNP Paribas Cardif 3](#_Toc497334126)

[3. Exploring the Data 4](#_Toc497334127)

[3.1 Structure 4](#_Toc497334128)

[3.2 Target imbalance 4](#_Toc497334129)

[3.3 Column names 4](#_Toc497334130)

[3.4 Empty rows 4](#_Toc497334131)

[4. Data Preparation 5](#_Toc497334132)

[4.1 Data cleaning 5](#_Toc497334133)

[5. Data exploration and feature selection 7](#_Toc497334134)

[5.1 Histograms 7](#_Toc497334135)

[5.2 Correlation analysis of numeric variables 8](#_Toc497334136)

[5.3 Zero and near-zero variance 8](#_Toc497334137)

[5.4 Variable screening by Information Value 9](#_Toc497334138)

[6. Logistic regression 11](#_Toc497334139)

[6.1 Logistic regression models 11](#_Toc497334140)

[6.2 Model evaluation 12](#_Toc497334141)

[7. Future directions 15](#_Toc497334142)

## 1. 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. 1

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.2 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.

## 2. Obtaining data: BNP Paribas Cardif

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 was 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 is evaluated by submission on the Kaggle website and scored by log loss ranking.

## 3. Exploring the Data

### 3.1 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.

### 3.2 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. For the purposes of this project, the data will be used as is, without resampling.

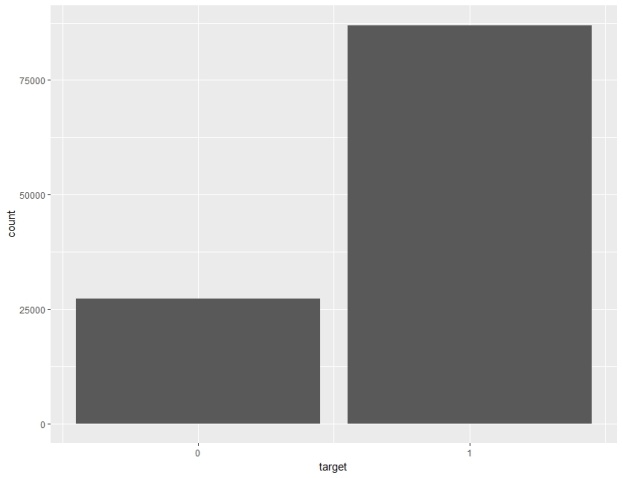


Figure : Count of 0 and 1 in dependent variable "target"

### 3.3 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.

### 3.4 Empty rows

The training data was analyzed for the presence of any empty rows, ie. rows with all values missing. None were found.

## 4. Data Preparation

Data preparation is an essential stage in data science that converts raw data into a format that can be used for analysis and model building. The results and accuracy of the model will be heavily influenced by the choices made at this phase. This process can be broadly divided into two steps:

1. Data cleaning

2. Feature engineering and selection

### 4.1 Data cleaning

Real-world data is typically unformatted, with inaccuracies, noise and inconsistencies. Data cleaning is therefore a critical step that gives the data the right format and quality for the analysis and subsequent modelling to be performed. This process includes but is not limited to the following3:

* Basics: eg. filtering outliers, removal of duplicates, renaming columns
* Sampling
* Data Partitioning
* Transformations
* Binning
* Data Replacement
* Attribute Weighting and Selection
* Attribute Generation
* Imputation of missing values

#### 4.1.0 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).

#### 4.1.1 Missing values

Upon first inspection of structure and summary, there appear to be many missing values (Figure 1). 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.

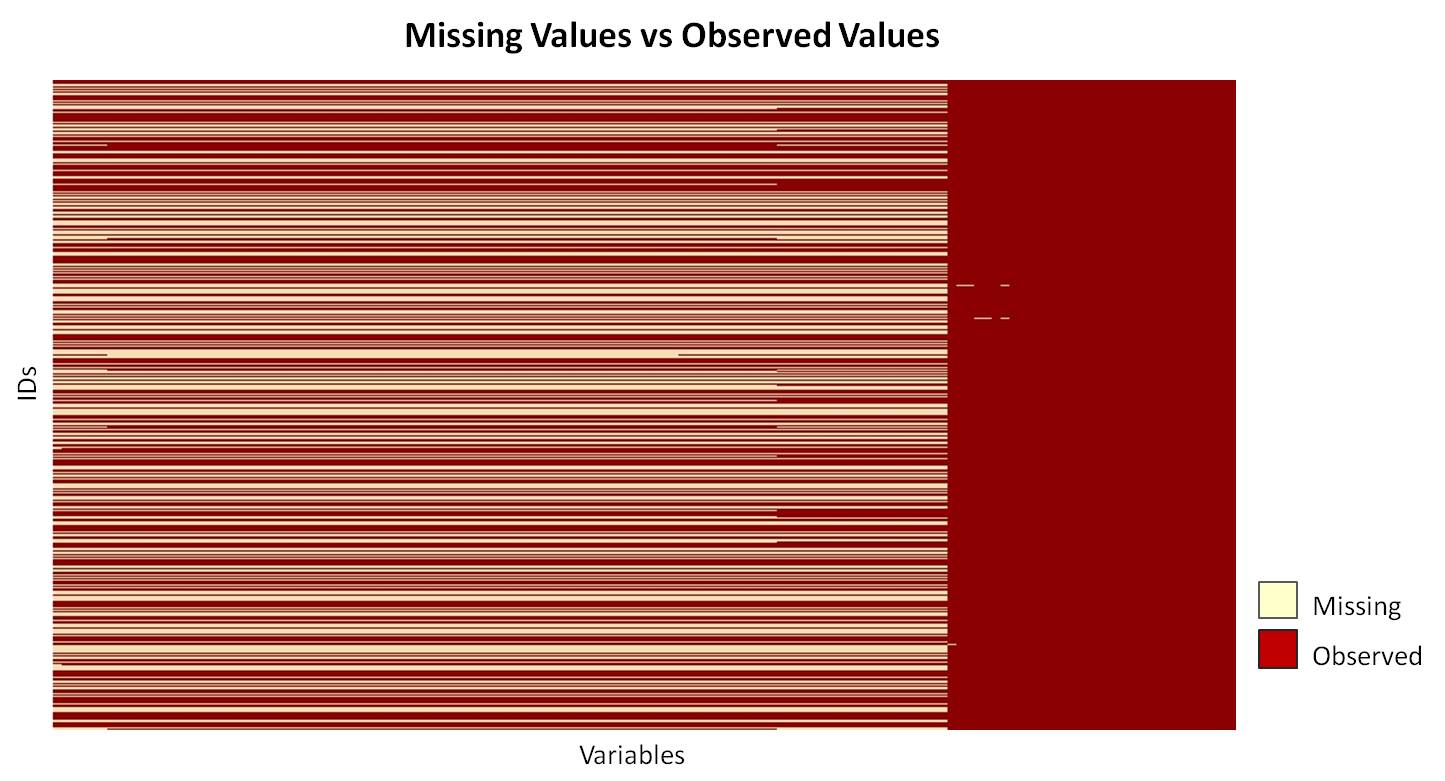


Figure : Missing Values vs Observed

#### 4.1.2 Separation of categorical and numeric variables

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

#### 4.1.3 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.

#### 4.1.4 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.

## 5. Data exploration and feature selection

In machine learning, feature selection is the process of selecting the most relevant attributes in the data for use in predictive model building. Choosing fewer, better features will enable better prediction performance with simpler models and less data. Typical selection methods are used to identify and remove superfluous, irrelevant or redundant attributes from the data that either do not contribute to model accuracy or even reduce accuracy.

### 5.1 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 (see section 5.3)

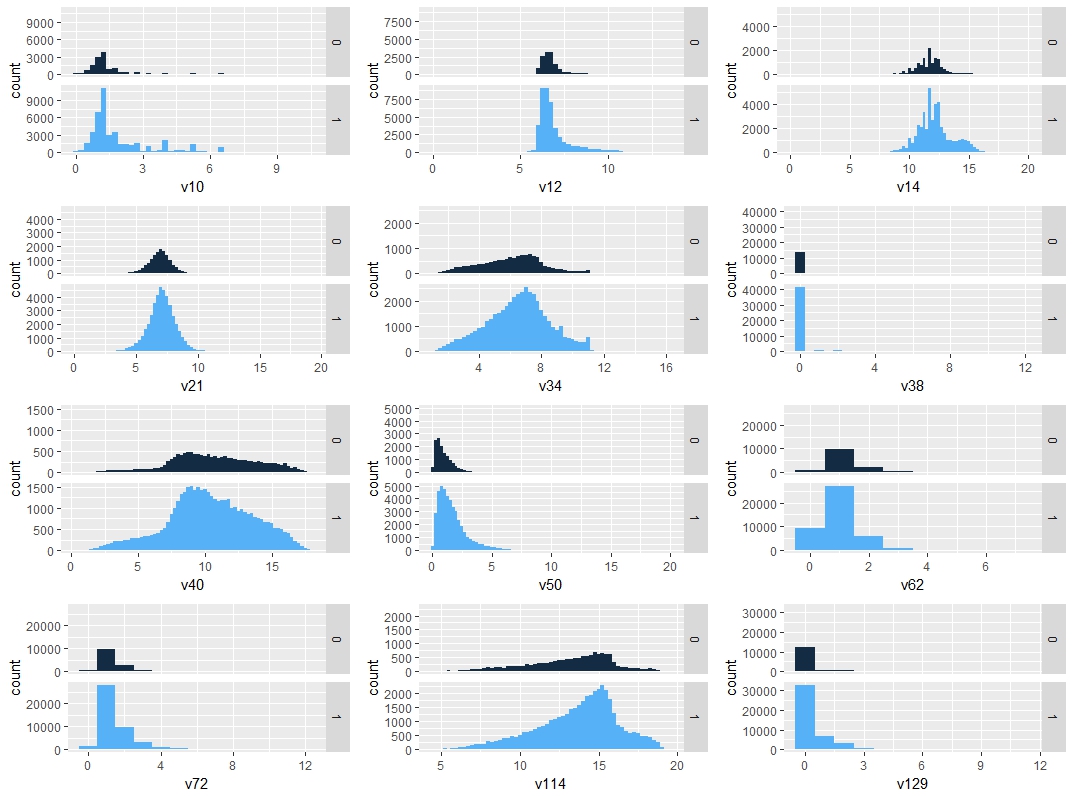
****

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

****

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

### 5.2 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.

### 5.3 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 |

### 5.4 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.4

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:5

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:6

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:

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 |

Highly correlated variables are shown highlighted in yellow.

## 6. Logistic regression

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, which is to identify claims suitable for accelerated approval in the BNP Paribas database.

Logistic regression generates the coefficients of a formula used to describe a logit transformation of the probability of an event occuring7:

Logistic regression equation

where p is probability of an event occurring, and the log-odds ratio is the natural logarithm of the ratio between the probability of an event occurring, to the probability that it will not occur:

Logit(p)=ln(p/(1-p))

Compared to linear regression, logistic regression does not choose parameters that minimize the sum of squared errors. Instead, parameters are chosen that maximize the likelihood of observing the sample values.

### 6.1 Logistic regression models

Logistic regression models were built and analyzed based on the exploration performed in section 5. Information value (IV) scores were used for variable ranking, but weight of evidence (WOE) was not used for imputation. By convention, only variables with medium or strong predictive power (IV 0.1-0.5) were used.

Table : Logistic regression model names, descriptions and variables

|  |  |  |
| --- | --- | --- |
| Model Name | Description | Variables |
| Log5Num | Numeric variables only, before correlation analysis | v50, v129, v62, v10, v14 |
| Log3Num | Numeric variables only, after correlation analysis | v50, v129, v62 |
| Log8NumCat | Numeric and categorical variables, before correlation analysis | v50, v31, v129, v10, v47, v62, v14, v66 |
| Log6NumCat | Numeric and categorical variables, after correlation analysis | v50, v31, v129, v47, v62, v66 |
| Log5aNumCat | Numeric and categorical variables, after correlation analysis | v50, v31, v129, v62, v66 |
| Log5bNumCat | Numeric and categorical variables, after correlation analysis | v50, v31, v129, v47, v66 |
| Log4NumCat | Numeric and categorical variables, after correlation analysis | v50, v31, v129, v66 |

As near-zero variance variables did not rank highly by IV, they were not considered for model building. Numeric variables demonstrating correlation were included to determine effect on model accuracy.

### 6.2 Model evaluation

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 were first evaluated by Area Under Curve (AUC) on the validation set prior to log loss ranking of test set predictions on the Kaggle website.

Table : Model AUC and Kaggle score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Variable Significance | AUC | Kaggle Score | AUC Rank | Kaggle Rank |
| Log5Num | \*\*\* v50, v129, v14  \* v10  . v62 | 0.69863 | 0.50538 | 7 | 6 |
| Log3Num | \*\*\* v50, v129  \*\* v62 | 0.69905 | 0.50545 | 6 | 7 |
| Log8NumCat | \*\*\* v50, v31B, v31C, v14, v66B, v66C  \*\* v129  v10, v47, v62 | 0.72705 | 0.48858 | 2 | 1 |
| Log6NumCat | \*\*\* v50, v31B, v31C, v66B, v66C  \*\* v129  v47, v62 | 0.72703 | 0.48864 | 3 | 3 |
| Log5aNumCat | \*\*\* v50, v31B, v31C, v129, v66B, v66C  \* v62 | 0.71939 | 0.49324 | 4 | 4 |
| Log5bNumCat | \*\*\* v50, v31B, v31C, v66B, v66C  \*\* v129  v47 | 0.72707 | 0.48860 | 1 | 2 |
| Log4NumCat | \*\*\* v50, v31, v129, v66 | 0.71937 | 0.49335 | 5 | 5 |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The receiving operator characteristic (ROC) curve is a fundamental tool for diagnostic model evaluation where the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity). Each point on the curve represents the sensitivity-specificity outcome corresponding to a specific threshold. The area under an ROC curve can be used as a measure of classifier performance independent of thresholding.

The Kaggle score was obtained by using the model to predict the probability of accelerated approval for each ID in the provided test set. The probabilities were submitted to the Kaggle competition site and ranked by logarithmic loss or log loss. Log loss quantifies accuracy by penalizing false predictions, therefore minimizing log loss is the desired outcome.

Based on AUC and Kaggle scoring, the top three models were, in order: Log5bNumCat, Log8NumCat and Log6NumCat when ranked by AUC, and Log8NumCat, Log5bNumCat and Log6NumCat when ranked by Kaggle score. All three scored very similarly by either metric, with less than 0.0001 differences between them. Because Log8NumCat was built using highly correlated variables, it was discarded in favour of the remaining two models, Log6NumCat and Log5bNumCat. These two models differ by a single variable, v62, which is included in Log6NumCat but not in Log5bNumCat

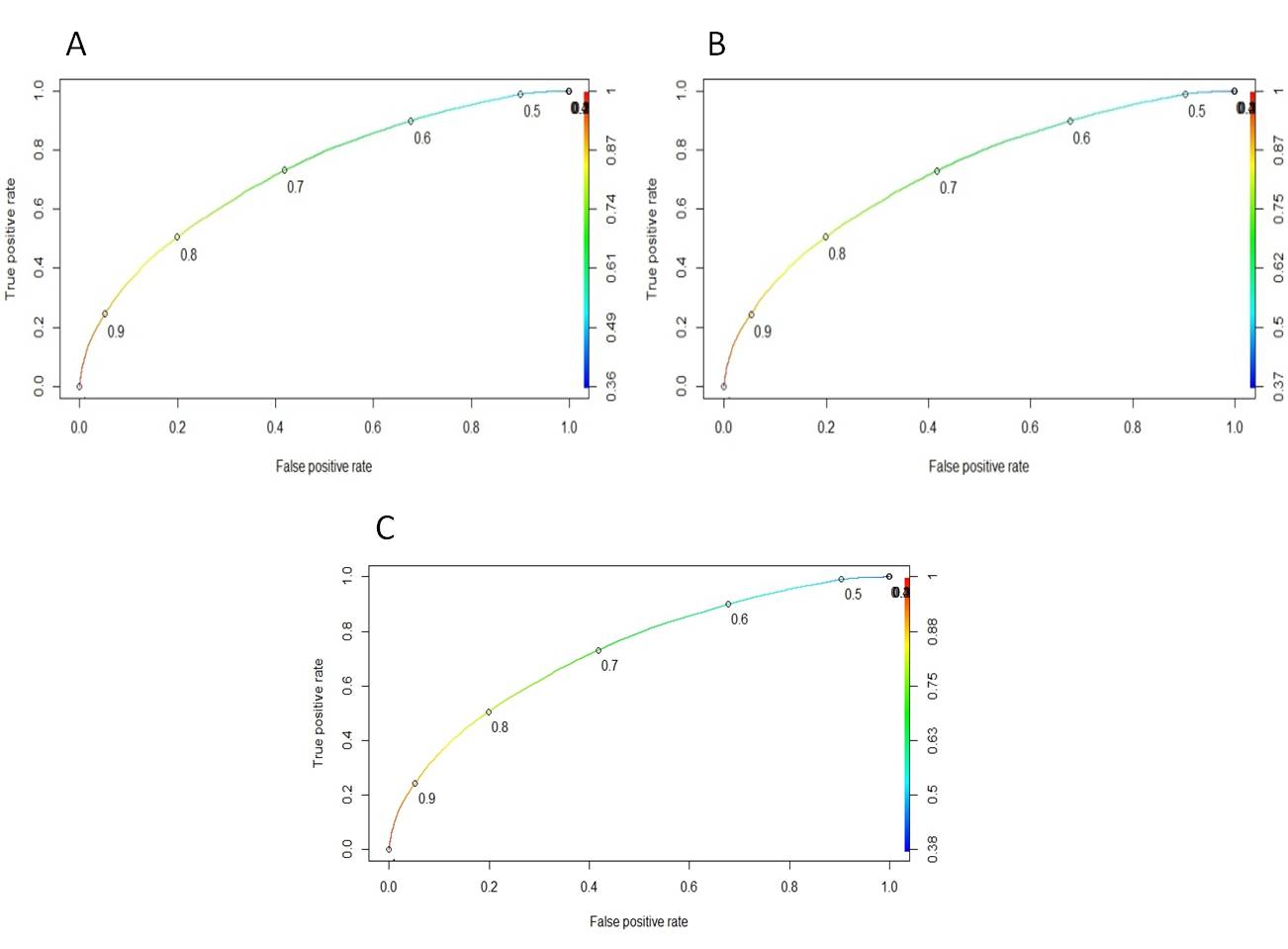


Figure : AUC for A. Log8NumCat, B. Log6NumCat and C. Log5bNumCat

A logistic regression model is said to have a better fit to the data if it can demonstrate a statistically significant improvement over a model with fewer predictors. As Log6NumCat was built with one more variable than Log5bNumCat, a likelihood ratio test was performed to determine if the observed difference was significant. Using the anova() function, a *p*-value of 0.1504 was obtained. Therefore, it was determined that the best model for predicting accelerated claims was Log5bNumCat, using just five predictors out of the original 131 variables in the available data – v50, v31, v129, v47 and v66.

The relative contribution to the model for each predictor is summarized in Table 9, while Figure 6 graphically demonstrates the effect on prediction for each variable and can be used to determine predicted probability of an accelerated claim, given a value for each predictor.

Table : Relative variable importance for Log5bNumCat

|  |  |
| --- | --- |
| Variable | Importance |
| v50 | 49.04988 |
| v31B | 6.087676 |
| v31C | 9.899482 |
| v129 | 2.992602 |
| v47B | 0.077762 |
| v47C | 0.34311 |
| v47D | 0.258011 |
| v47E | 0.26282 |
| v47F | 0.190158 |
| v47G | 0.160522 |
| v47H | 0.109408 |
| v47I | 0.129186 |
| v47J | 0.12598 |
| v66B | 11.18959 |
| v66C | 25.4425 |

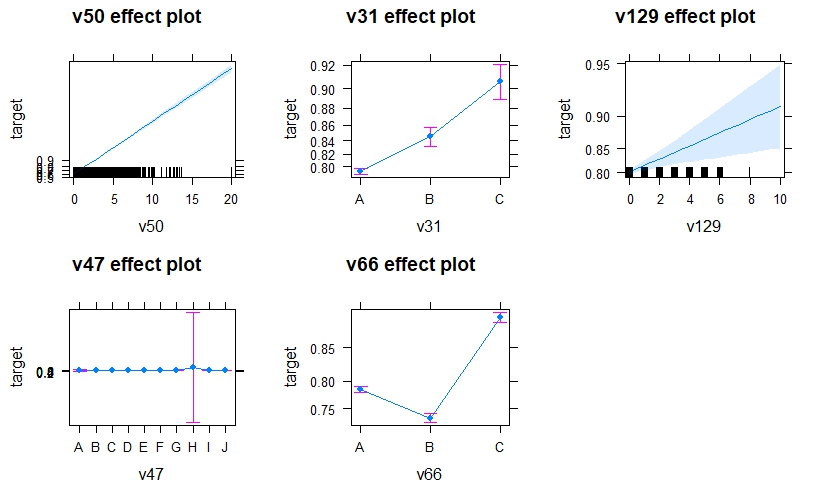


Figure : Graphing predicted values using model Log5bNumCat

## 7. Future directions

With additional time and to further improve the accuracy of the model, there are various techniques that would be of interest to pursue. As described in section 3, the data displays a modest 3:1 imbalance in target ratio. This could be addressed by rebalancing the training data by oversampling the minority class, undersampling the majority class or creating new minority classes from the majority class.

While the current model was assessed on both the validation data set and the test set, this could be further refined by performing k-fold cross-validation on the training data provided instead of using a single partition into training and validation sets. Because of the obvious patterns of missing values across the columns, it could be valuable to explore the predictor capability of the missing values in those variables.

Lastly, it would be very interesting to compare the outcome and accuracy of these logistic regression models to additional models built using decision tree methods, such as CART and random forest.

## 8. References

1. Baxter, Lawrence. The social value of insurance. *Policy and Public Affairs, Chartered Insuranc e Institute*. Accessed Oct 17, 2017. http://www.cii.co.uk/knowledge/policy-and-public-affairs/articles/the-social-value-of-insurance/32307
2. Patrick, Bart. Insurance: Improving claims management. *The Actuary*. Accessed Oct 17, 2017. http://www.theactuary.com/archive/old-articles/part-6/insurance-3A-improving-claims-management/
3. Wähner, Kai. Data Preprocessing vs. Data Wrangling in Machine Learning Projects. Accessed Oct 24, 2017 https://www.infoq.com/articles/ml-data-processing
4. Larsen, Kim. Data exploration with Weight of evidence and Information Value. Accessed Oct 24, 2017. http://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/
5. Upadhyay, Roopam. Information Value (IV) and Weight of Evidence (WOE) – A Case Study from Banking (Part 4). Accessed Oct 25, 2017. http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/
6. Bhalla, Deepanshu. Weight of Evidence (WOE) and Information Value explained. Accessed Oct 25, 2017. http://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html
7. MedCalc. Logistic Regression. Accessed Oct 31, 2017. https://www.medcalc.org/manual/logistic\_regression.php