Capstone Report

Monisha Gopal

November 2017

1 Introduction

BNP Paribas Cardif is an global insurance company that specializes in personal insurance. Not only do they have to deal with an increasing number of claims, their customers are also expecting them to handle claims as fast as possible. Usually claims go through a number of checks before being approved. However, BNP Paribas Cardif hopes to speed up that process using data science.

2 Dataset

BNP Paribas Cardif provided an anonymized dataset on kaggle.com with two categories of claims described on kaggle.com as follows:

1. claims for which approval could be accelerated leading to faster payments
2. claims for which additional information is required before approval

The goal is to predict the category of a claim based on information available early in the claims process.

Link: https://www.kaggle.com/c/bnp-paribas-cardif-claims-management

3 Initial Look

Three files are provided:

1. train.csv - training set with target (dependent variable)
2. test.csv - test set without target
3. samplesubmission.csv - sample submission with correct format

The dataset contains 133 features named 'ID', 'target', and 'v1' through 'v131'. There are both categorical and numerical features. None of the categorical features are ordinal (specified on kaggle).

The main limitation of this dataset is that the features are anonymized. Because of this, we can't use domain knowledge to eliminate features or predict the distributions of certain features.

4 Data Wrangling

Main steps for data wrangling:

1. Remove all features that had more than 25 percent NA values
    - Went from 131 features to 29 features
2. Remove the outliers in continuous variables and replaced with NA values
3. Replace the NA values in continuous variables with the mean
4. Replace the NA values in categorical variables with the mode
5. Remove features with zero/near-zero variance
    - Went from 29 features to 26 features
6. Remove categorical variables with more than 20 categories
    - Went from 26 features to 22 features
7. Remove features with perfect multicollinearity
    - Went from 22 features to 19 features
8. Remove features that were highly correlated with others using VIF
    - Went from 19 features to 14 features

The features remaining include, v10, v12, v21, v24, v47, v50, v52, v62, v66, v71, v72, v91, v114, v129.

5 Feature Enrichment

The main methods for feature enrichment used were Weight of Evidence (WoE) and Information Value (IV).

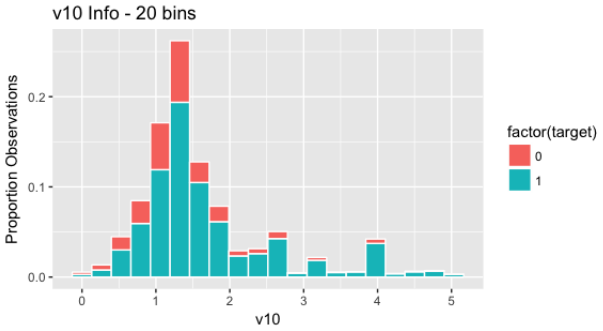
For continuous variables, initially we take the range and divide it up into 20 equally spaced bins. We calculate the weight of evidence for each of those bins and graph WOE against the bins, also including information about what percentage of total observations are contained in each bin. The rule of thumb is that every bin should contain at least 5 percent of the total observations. So from the original bins, adjacent bins were combined if either they had similar WOE scores or if they both contained fewer than 5 percent of the observations.

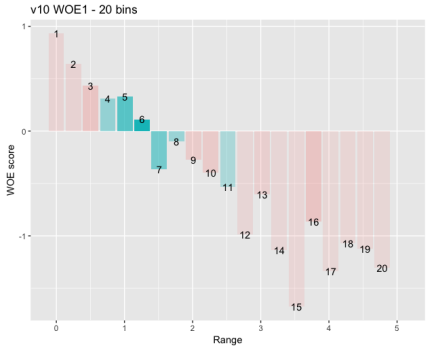
Similarly for continuous variables, we assign each category as a bin. If categories have similar WOE scores, they are combined.

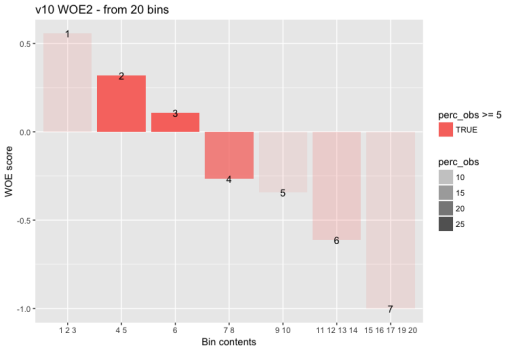
After WOE scores are calculated, the information value for each feature is also calculated. The features are then ranked by significance (predictive power). Features with extremely low predictive power are tossed, and the rest have their values replaced with their respective WOE scores.

5.1 v10 - Continuous feature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.000001	1.050328	1.312910	1.615455	1.838074	5.032824

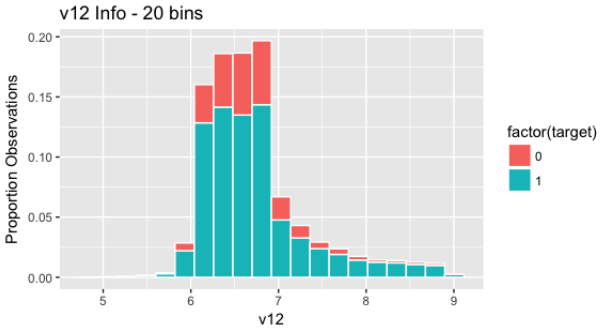


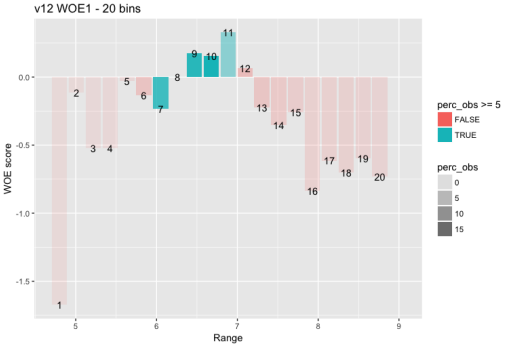


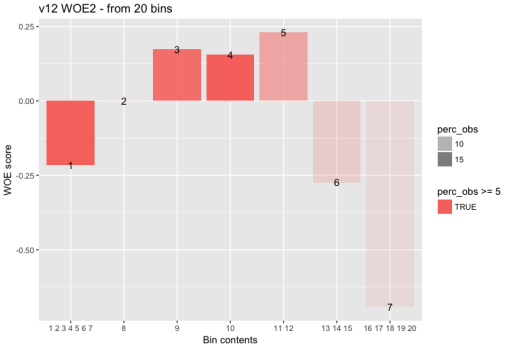


5.2 v12 - Continuous feature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4.799	-6.324	-6.615	-6.725	-6.893	-8.972

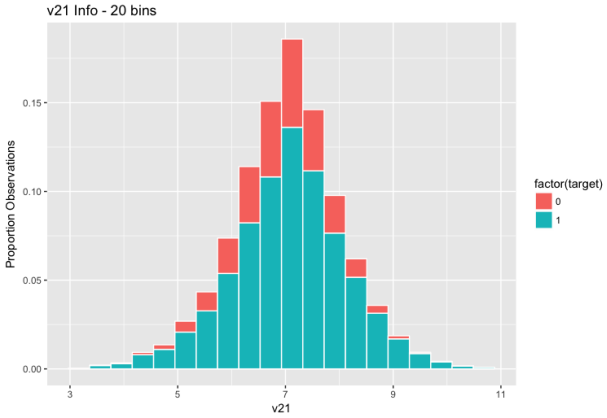


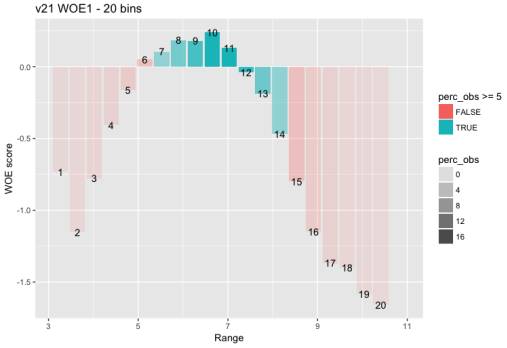


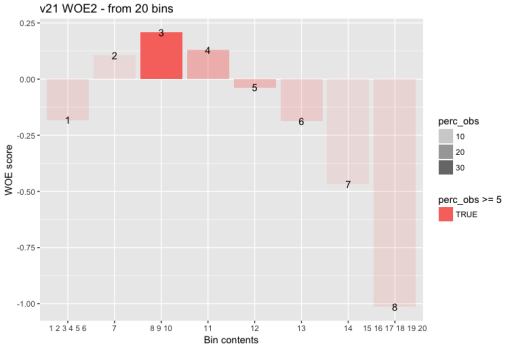


5.3 v21 - Continuous feature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.268	-6.423	-7.035	-7.032	-7.661	-10.788

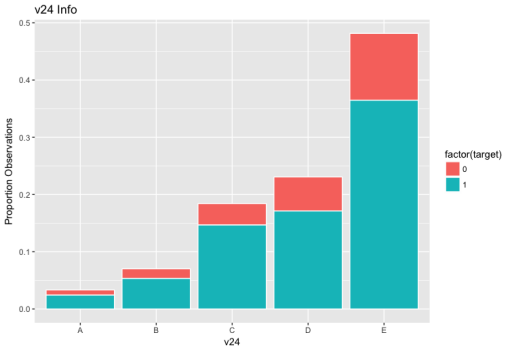


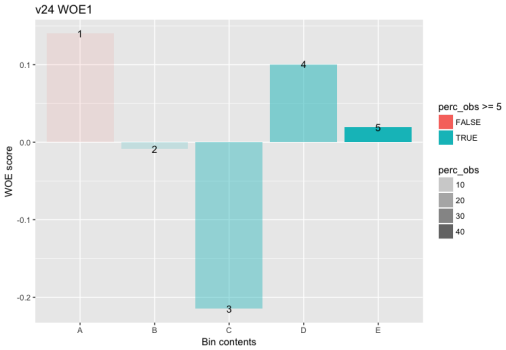




5.4 v24 - Categorical feature (TOSSED)

A	B	C	D	E
1904	4014	10528	13190	27525

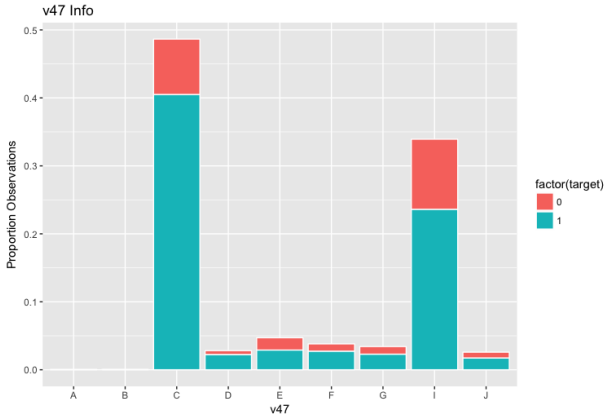


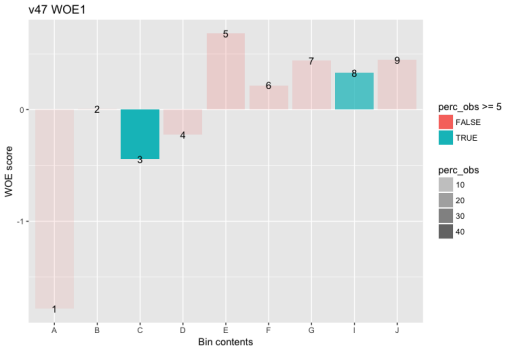


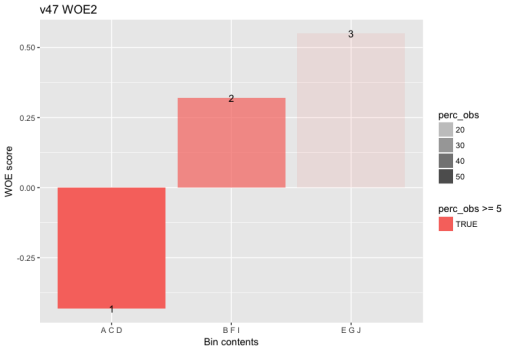
The IV for the initial binning of v24 was 0.01096248 which is considered not useful for predictions so this variable was tossed.

5.5 v47 - Categorical feature

A	B	C	D	E	F	G	I	J
20	25	27811	1609	2693	2177	1953	19392	1481

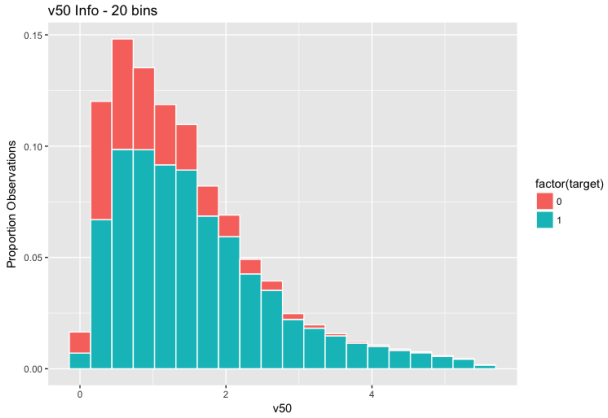


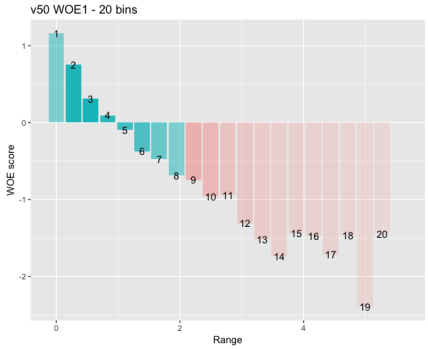


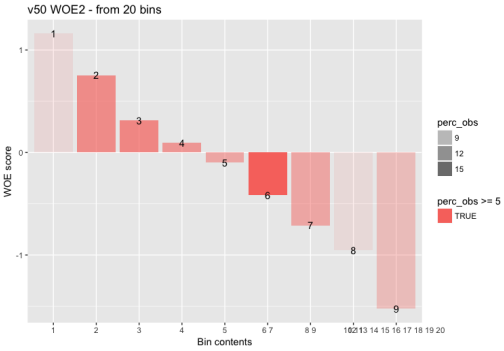


5.6 v50 - Continuous feature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.000001	0.658793	1.216345	1.460364	1.971976	5.549116

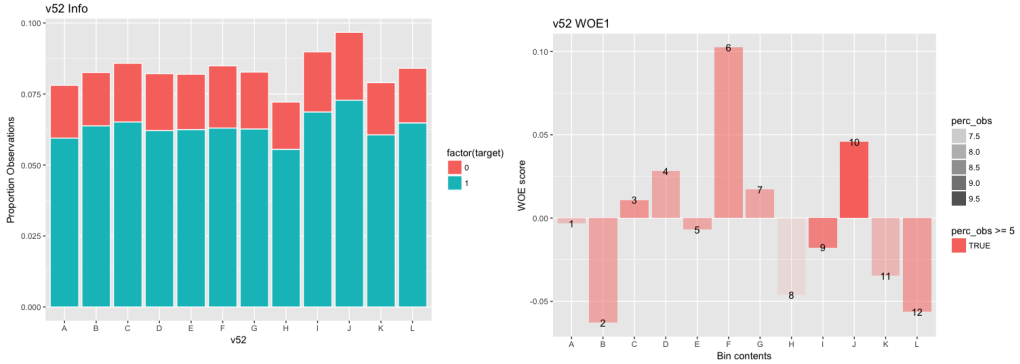






5.7 v52 - Categorical feature (TOSSED)

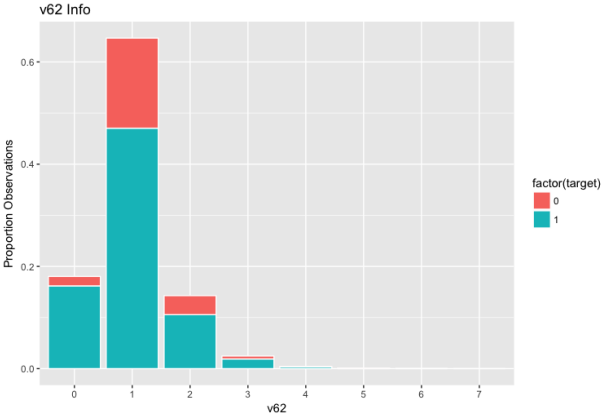
A	B	C	D	E	F	G	H	I	J	K	L
4462	4720	4905	4697	4685	4854	4728	4125	5136	5529	4515	4805

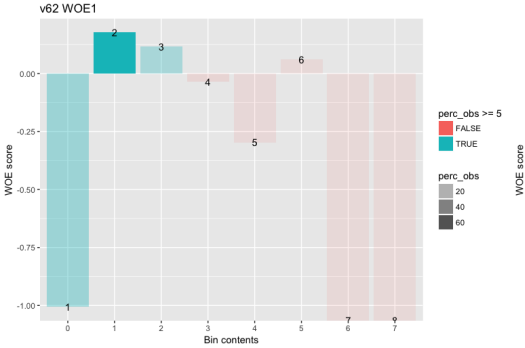


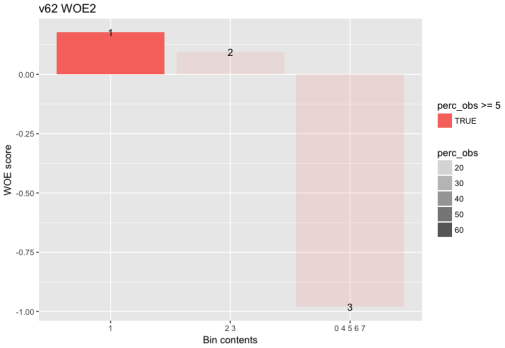
The IV for the initial binning of v52 was 0.002302835 which is considered not useful for predictions so this variable was tossed.

5.8 v62 - Categorical feature

0	1	2	3	4	5	6	7
10314	36965	8161	1412	238	48	22	1

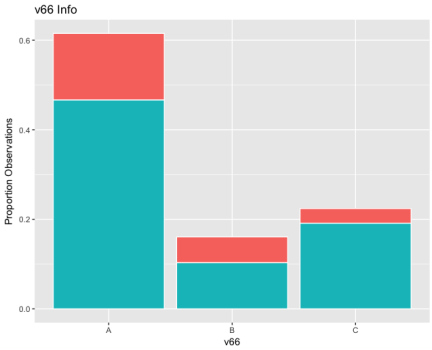


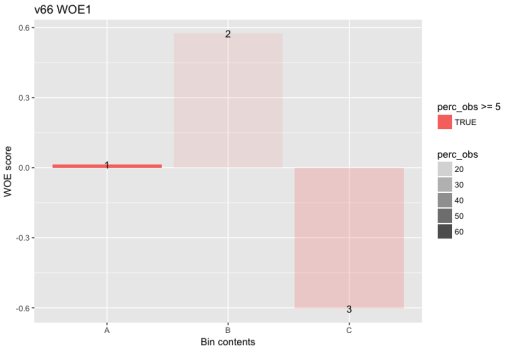




5.9 v66 - Categorical feature

A	B	C
35164	9188	12809





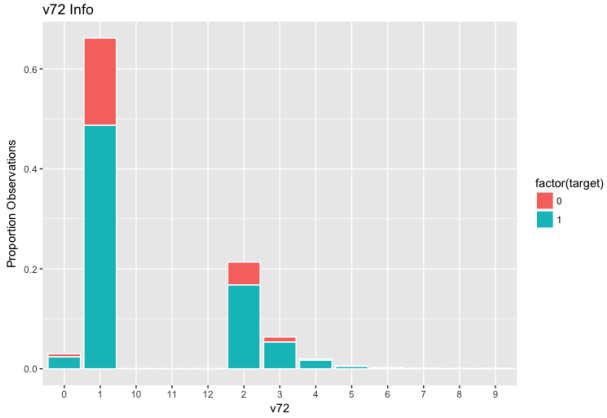
5.10 v71 - Categorical feature (TOSSED)

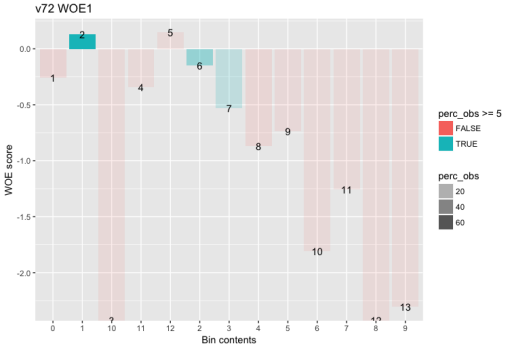
A	B	C	D	F	G	I	K	L
0	15111	4516	1	37519	3	10	1	0

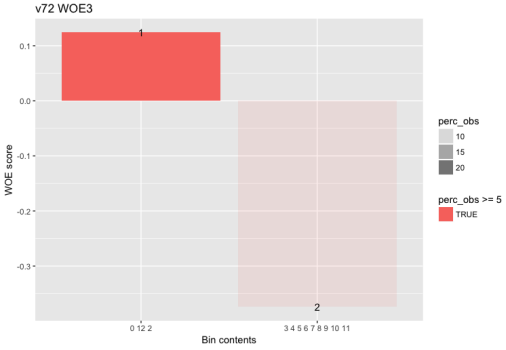
The IV for the binning with no empty bins of v71 was 0.0059342 which is considered not useful for predictions so this variable was tossed.

5.11 v72 - Categorical feature

0	1	10	11	12	2	3	4	5	6	7	8	9
1708	37824	30	11	15	12197	3635	1129	329	143	61	46	33

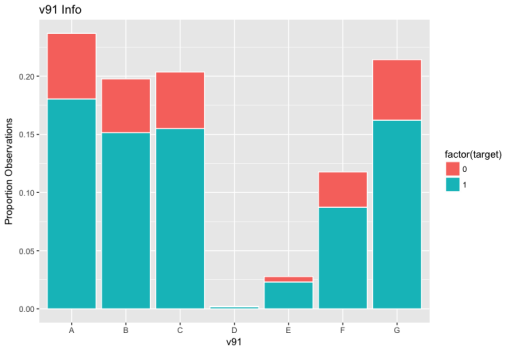


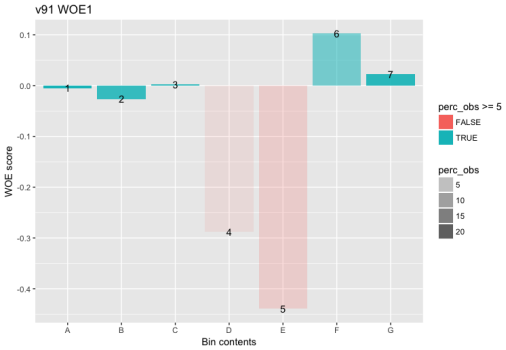




5.12 v91 - Categorical feature (TOSSED)

A	B	C	D	E	F	G
13534	11300	11638	126	1588	6729	12246

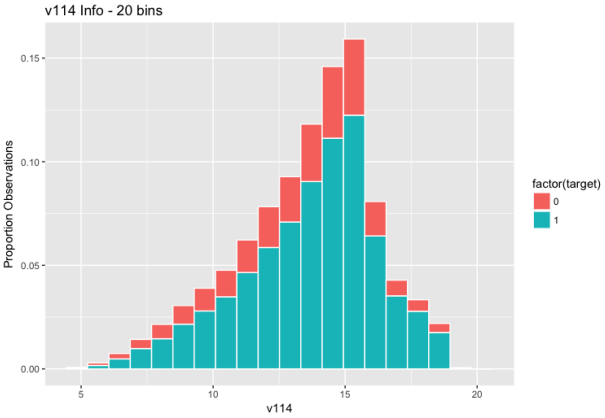


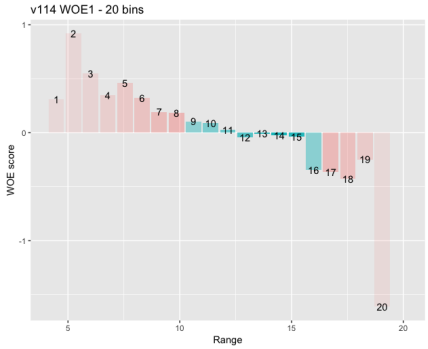


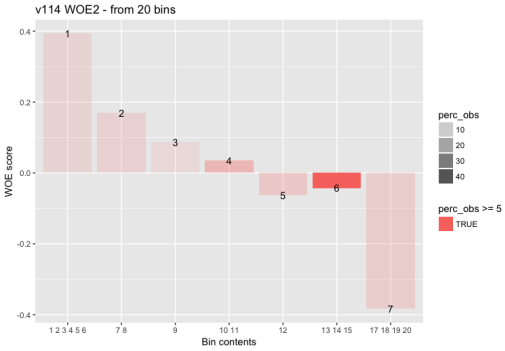
The IV for the binning with no empty bins of v91 was 0.006104932 which is considered not useful for predictions so this variable was tossed.

5.13 v114 - Continuous feature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.479	11.981	14.040	13.572	15.371	19.820

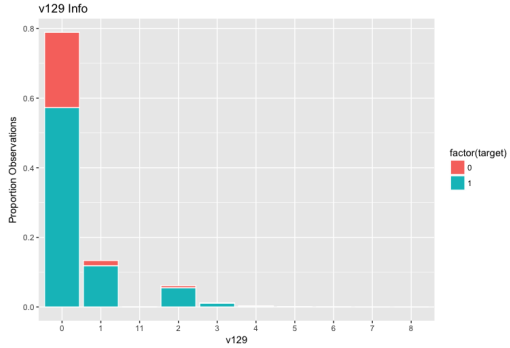


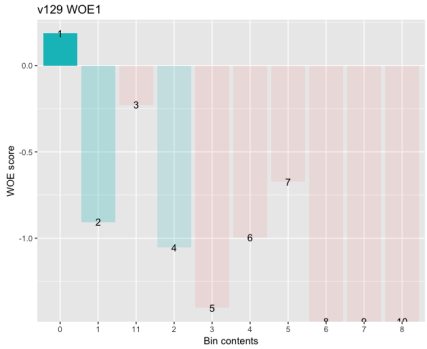


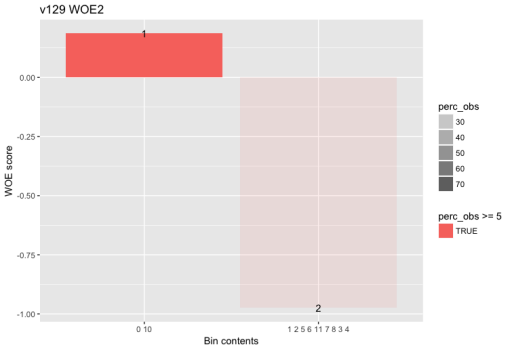


5.14 v129 - Categorical feature

0	1	11	2	3	4	5	6	7	8
45103	7630	5	3499	685	183	29	25	1	1







5.15 Information Value Summary

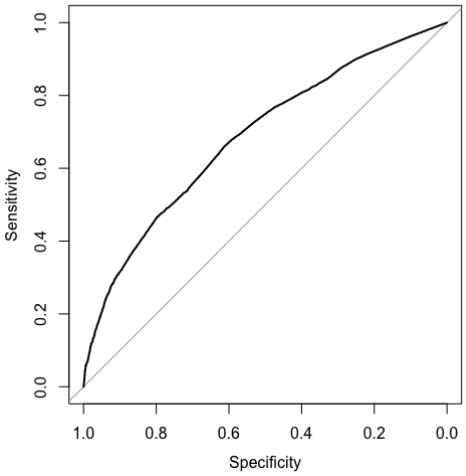
Variable	IV	Significance
v10	0.1461525	**
v12	0.05399145	*
v21	0.08575021	*
v24	0.01096248	TOSS
v47	0.1573225	**
v50	0.4500647	***
v52	0.002302835	TOSS
v62	0.146551	**
v66	0.1282981	**
v71	0.00593427	TOSS
v72	0.04107886	*
v91	0.006104932	TOSS
v114	0.03501795	*
v129	0.175497	**

6 Machine Learning

Since this is a binary classification problem, the main machine learning model used was logistic regression. Three models were created, one using the top 5 most significant variables by IV score, one using the top 8 most significant variables, and one using the top 10 most significant variables.

6.1 Model 1: Top 5 Most Significant Variables

Variable	Coefficient
Intercept	0.9903
v50	-1.0215
v129	-0.3731
v47	-0.6610
v10	-0.0854
v62	-0.1588



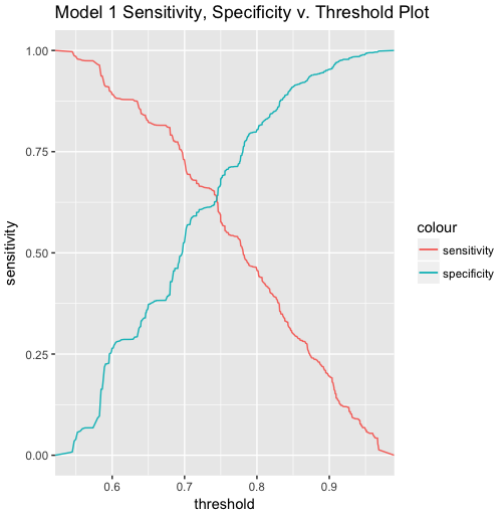


Figure 1: ROC Curve Model 1

	Reference	
	FALSE	TRUE
0	8621	15813
1	5029	27697

- Threshold = 0.7444081
- ROAUC = 0.6856

- Specificity = 0.6316
- Sensitivity = 0.6366
- Positive Predictive Power = 0.8463
- Negative Predictive Power = 0.3528
- Accuracy = 0.6354

6.2 Model 2: Top 8 Most Significant Variables

Variable	Coefficient
Intercept	0.99449
v50	-0.99129
v129	-0.35418
v47	-0.62245
v10	-0.07636
v62	-0.04564
v66	-0.81682
v21	-0.22871
v12	-0.10436

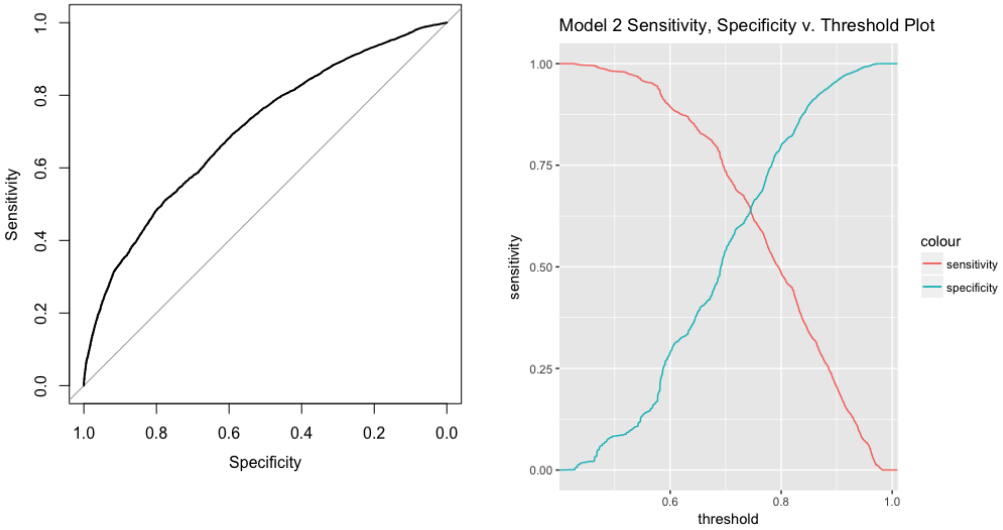


Figure 2: ROC Curve Model 2

	Reference	
	FALSE	TRUE
0	8621	15813
1	5029	27697

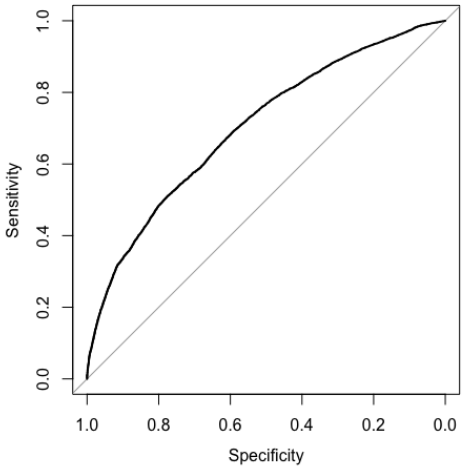
- Threshold = 0.7451645
- ROAUC = 0.7022
- Specificity = 0.6396
- Sensitivity = 0.6399
- Positive Predictive Power = 0.8498
- Negative Predictive Power = 0.2388
- Accuracy = 0.6398

6.3 Model 3: Top 10 Most Significant Variables

Variable	Coefficient
Intercept	0.9755912
v50	-0.9916858
v129	-0.3632729
v47	-0.6267445
v10	-0.0761522
v62	-0.0498536
v66	-0.8176888
v21	-0.2268336
v12	-0.1032023
v72	0.0265222
v114	0.0008738

	Reference	
	FALSE	TRUE
0	8741	15641
1	4909	27869

- Threshold = 0.7450126
- ROAUC = 0.7023
- Specificity = 0.6404
- Sensitivity = 0.6405
- Positive Predictive Power = 0.8502
- Negative Predictive Power = 0.3585
- Accuracy = 0.6405



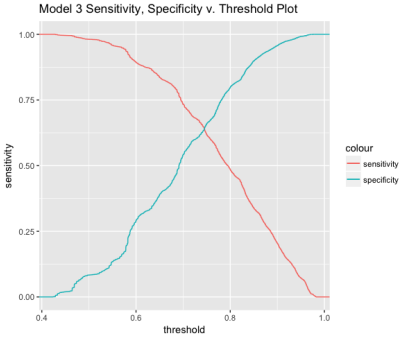


Figure 3: ROC Curve Model 3

6.4 Kaggle Results

Submission and Description	Private Score	Public Score
model3.csv 5 minutes ago by Monisha Gopal Top 10 variables	0.50107	0.50170
model2.csv 6 minutes ago by Monisha Gopal Top 8 variables	0.50112	0.50169
model1.csv 6 minutes ago by Monisha Gopal Top 5 variables	0.50927	0.50998

Figure 4: Kaggle Private and Public Results

7 Conclusion

Based on the results, model 2 has the best performance of the 3 models. Adding more predictors in model 3 did not give noticeably better accuracy, specificity, sensitivity, etc, than that of model 2. Model 2 also performed better on general datasets based on Kaggle results.

BNP Paribas Cardif Claims Management could use this model as a way to quickly classify which claims can definitely go through an accelerated process and which need more review. That way, they can satisfy their customers need for speed without losing too much revenue.