Comparison of Resampling Techniques for Treatment of Unbalanced Data in Predictive Modelling : Loan Default Prediction

Case study by Data Science Nigeria

Kenneth Low Yan Wei
Peace Tay Jiunn Ching
Yap Pin Yaw
Supervisor: Prof. Kam Tin Seong

SMU SINGAPORE MANAGEMENT UNIVERSITY

1. Introduction

Background and Objective

Background

Financial institutions often use credit risk classification models to identify the risk of borrowers, to make informed business decisions. However, such data sets are often highly unbalanced, which can have serious negative effects on the classification performance of predictive algorithms. This is because traditional machine learning models and evaluation metrics assume a balanced data distribution.

Problem statement

There had been many proposed techniques in dealing with classification of unbalanced datasets, one of which is adopting resampling techniques to artificially rebalance binary classification datasets. However, the performance of using the various resampling techniques (i.e Weighting, Oversampling, Tomek, and SMOTE) using the various predictive modelling algorithms can still be improved.

Objective

In this paper, we present a study to identify combinations of resampling methods and predictive models will produce the best performance. The combination of the best performing resampling type and predictive algorithm can be used to produce a better predictive model, as well as address the problem of unbalanced data.

Data Preparation

Remove Variables

Remove variables that have >50% missing values as well as data that are irrelevant to the customer (longitude & latitude) to prevent biasness in the models.

Remove highly correlated variables

Remove variables that have high correlation for performing predictive modelling (e.g. Total due previous loans and loan amount of previous loans)

Grouping variables

Group categories in variables that have relatively low count (e.g. Student, and retired employment labels)

Data Transformation

Transformed variables into binary format to serve better for predictive modelling (e.g. Presence of referral, late payment indication)

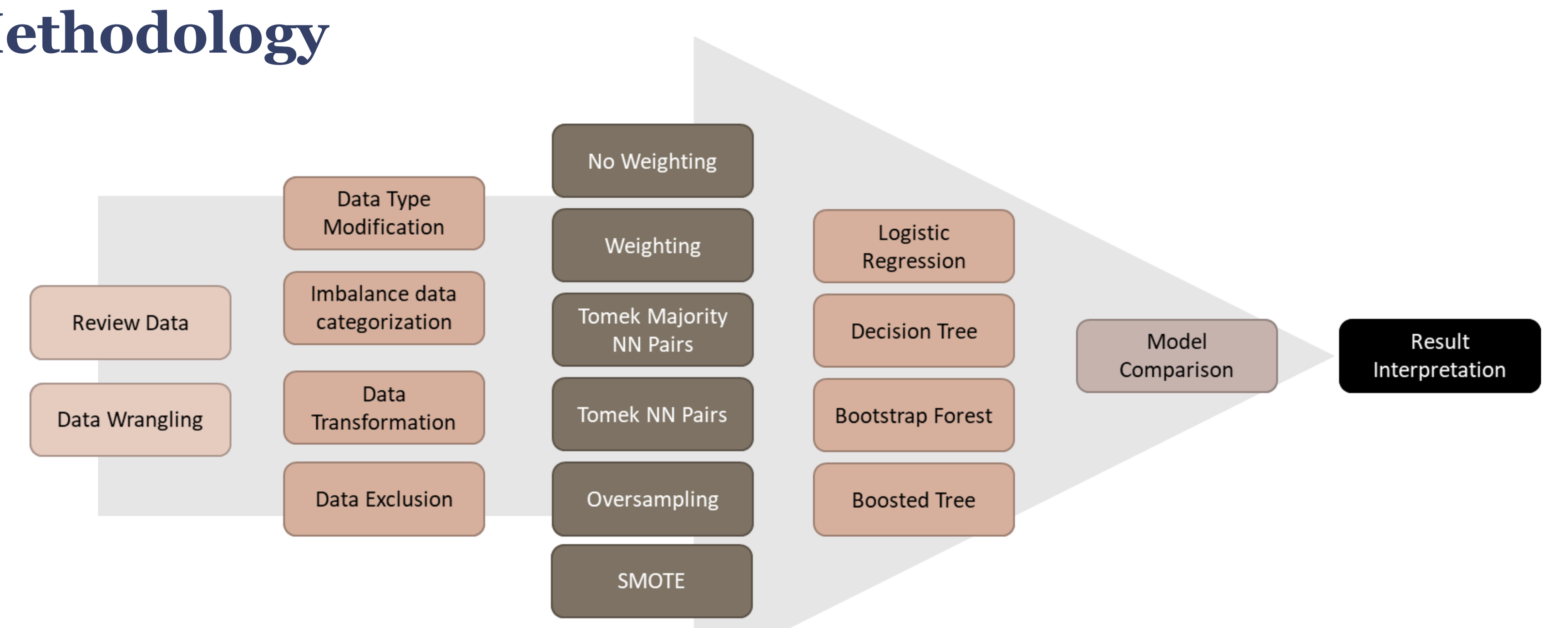
Join Data Sets

Demographic data

Current loan data

Previous loan data

Methodology



Overview

- The dataset contains loan and demographic data for the period of 2016-2017
- Each loan has been mapped as either default or no default
- The data used consists of a binary target and 12 independent variables

Predictive modelling

- SAS JMP Pro 16 was used
- With 5 resampling methods and 4 predictive algorithm used, a total of 20 predictive models was developed
- The 4 predictive modelling algorithm used are: (1) Logistic Regression, (2) Decision Tree, (3) Bootstrap Forest, (4) Boosted Tree

Predictive definition

		True Class	
		No Default	Defaults
Predicted Class	No Default	True Positive	False Positive
	Defaults	False Negative	True Negative

Resampling Techniques

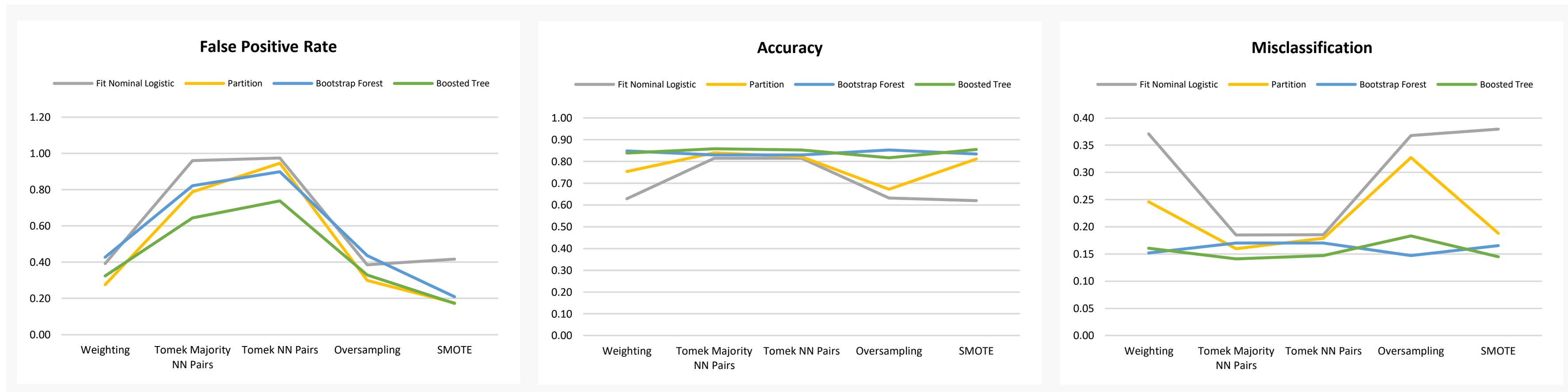
Resampling Technique	Description	Formula
No Weighting	No weighting allocated to under-represented class. To use as baseline comparison	-
Weighting	To use a frequency column that assigns a weight of 1 to majority cases and the ratio of number of majority / number of minority to the minority cases	$\text{If}\begin{pmatrix} good\_bad\_flag == \text{"Good"} \Rightarrow 1 \\ \text{else} \Rightarrow \frac{11165}{2528} \end{pmatrix}$
Over Sampling	To use a frequency column that assigns a weight of 1 to majority cases and a non-zero random integer to the minority cases.	$\text{If}\begin{pmatrix} good\_bad\_flag == \text{"Good"} \Rightarrow 1 \\ \text{else} \Rightarrow \text{Round}\left(\text{Random Normal}\left(\frac{11165}{2528}, 1\right), 0\right) \end{pmatrix}$
Tomek Links (Tomek Majority NN Pairs) Tomek Links (Tomek NN Pairs)	A Tomek Link is a pair of nearest neighbours that fall into different classes. Tomek links attempts to better define the boundary between the minority and majority classes by removing observations from the majority class that are "close" to minority class observations to better define cluster borders	JMP Imbalanced Classification Add in
SMOTE	Generates new data observations that are similar to the existing minority class observations rather than replicating them using the Gower distance and performing K – Nearest Neighbors on the minority class	JMP Imbalanced Classification Add in

2. Model Assessment

Table of Evaluation Metrics

	Sampling Types	Models	Misclassification Rate	Precision	Accuracy	Recall	False Positive Rate
			Test				
1	No Weighting	Logistic Regression	0.19	0.82	0.81	0.99	0.98
2		Decision Tree	0.16	0.85	0.84	0.98	0.78
3		Bootstrap Forest	0.17	0.83	0.83	0.99	0.87
4		Boosted Tree	0.15	0.86	0.85	0.98	0.72
5	Weighting	Logistic Regression	0.37	0.88	0.63	0.63	0.39
6		Decision Tree	0.25	0.92	0.75	0.76	0.28
7		Bootstrap Forest	0.15	0.90	0.85	0.91	0.43
8		Boosted Tree	0.16	0.92	0.84	0.88	0.32
9	Oversampling	Logistic Regression	0.37	0.88	0.63	0.64	0.39
10		Decision Tree	0.33	0.91	0.67	0.67	0.30
11		Bootstrap Forest	0.15	0.90	0.85	0.92	0.44
12		Boosted Tree	0.18	0.92	0.82	0.85	0.33
13	Tomek Majority NN Pairs	Logistic Regression	0.19	0.82	0.81	0.99	0.96
14		Decision Tree	0.16	0.85	0.84	0.98	0.79
15		Bootstrap Forest	0.17	0.84	0.83	0.98	0.82
16		Boosted Tree	0.14	0.87	0.86	0.97	0.64
17	Tomek NN Pairs	Logistic Regression	0.19	0.82	0.81	0.99	0.97
18		Decision Tree	0.18	0.82	0.82	0.996	0.95
19		Bootstrap Forest	0.17	0.83	0.83	0.995	0.90
20		Boosted Tree	0.15	0.85	0.85	0.99	0.74
21	SMOTE	Logistic Regression	0.38	0.63	0.62	0.65	0.42
22		Decision Tree	0.19	0.83	0.81	0.801	0.18
23		Bootstrap Forest	0.17	0.82	0.83	0.874	0.21
24		Boosted Tree	0.14	0.85	0.86	0.88	0.17
25		Optimised Boosted Tree	0.14	0.86	0.86	0.84	0.12

Exploratory Data Analysis



- ✓ We observe Tomek Majority NN Pairs & Tomek NN Pairs resampling technique gives the poorest performance when evaluating False Positive Rates. This is undesirable for loan default prediction
- ✓ We observe that Tomek Majority NN Pairs and Tomek NN Pairs resampling technique gives the highest performance when evaluation Recall rates
- ✓ We observe that SMOTE resampling technique provides low False Positive Rates

Model Comparison

Some assumptions were made about money lending companies are that they:

1 **Are profit driven** + 2 **Reduce loss due to default** = **Low False Positive Rate**

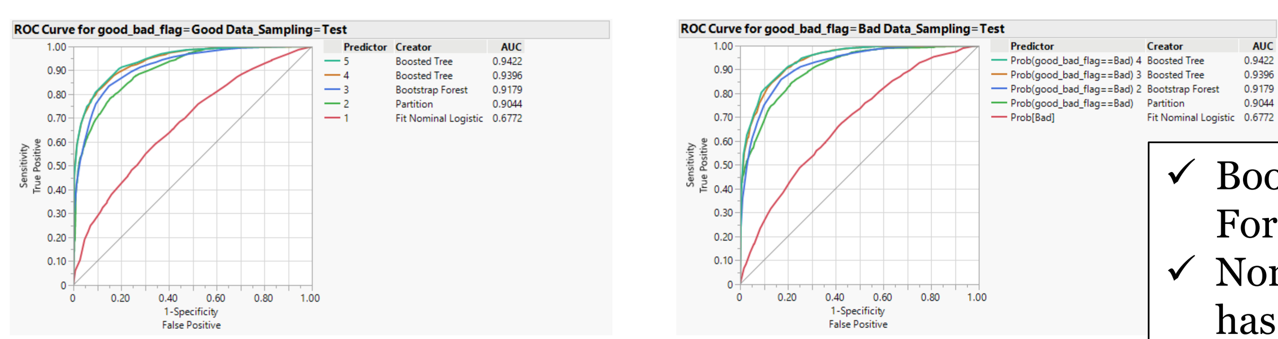
The efficacy of the models is evaluated based on: -

1 **Low False Positive** 2 **High Accuracy** 3 **Low Misclassification Rate**

Based on the results in the Table of Evaluation Metrics, the best result was given by using the **SMOTE** resampling method.

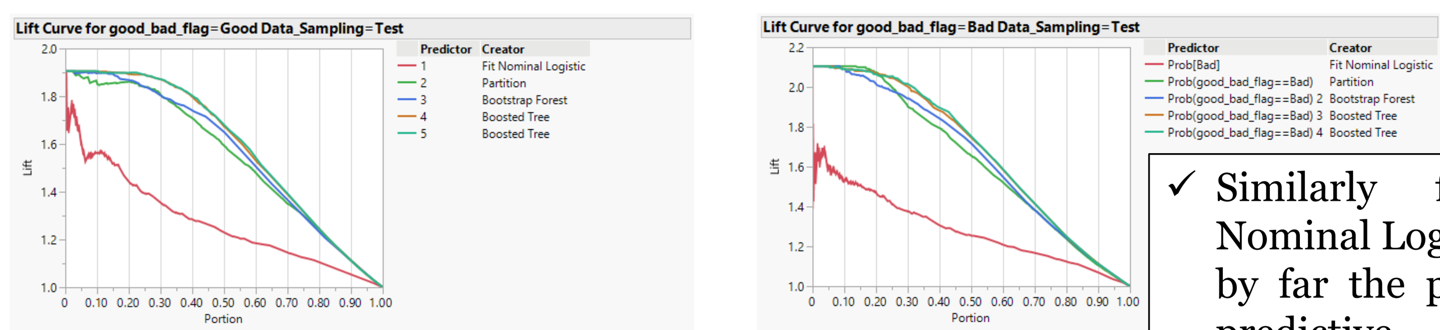
3. Evaluation and Analysis

ROC Curve for SMOTE



- ✓ Boosted Tree and Bootstrap Forest have the Highest AUC
- ✓ Nominal Logistic Regression has ~30% lower AUC as compared to Boosted Tree

Lift Curve for SMOTE



- ✓ Similarly for Lift Curve, Nominal Logistics Regression is by far the poorest performing predictive algorithm when using SMOTE

AUC Comparison for SMOTE

AUC Comparison for good_bad_flag=Bad for Data_Sampling=Test

Predictor	AUC	Std Error	Lower 95%	Upper 95%
Prob[Bad]	0.6772	0.0066	0.6641	0.6900
Prob(good_bad_flag==Bad)	0.9044	0.0036	0.8971	0.9111
Prob(good_bad_flag==Bad) 2	0.9179	0.0033	0.9111	0.9241
Prob(good_bad_flag==Bad) 3	0.9396	0.0027	0.9341	0.9446
Prob(good_bad_flag==Bad) 4	0.9422	0.0026	0.9368	0.9471

Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob>ChiSq
Prob[Bad]	Prob(good_bad_flag==Bad)	-0.227	0.0068	-0.240	-0.214	1129.3	<.0001*
Prob[Bad]	Prob(good_bad_flag==Bad) 2	-0.241	0.0061	-0.253	-0.229	1555.1	<.0001*
Prob[Bad]	Prob(good_bad_flag==Bad) 3	-0.262	0.0063	-0.275	-0.250	1754.6	<.0001*
Prob[Bad]	Prob(good_bad_flag==Bad) 4	-0.265	0.0063	-0.277	-0.253	1751.2	<.0001*
Prob(good_bad_flag==Bad)	Prob(good_bad_flag==Bad) 2	-0.013	0.0028	-0.019	-0.008	23.262	<.0001*
Prob(good_bad_flag==Bad)	Prob(good_bad_flag==Bad) 3	-0.035	0.0031	-0.041	-0.029	125.71	<.0001*
Prob(good_bad_flag==Bad)	Prob(good_bad_flag==Bad) 4	-0.038	0.0033	-0.044	-0.031	131.83	<.0001*
Prob(good_bad_flag==Bad) 2	Prob(good_bad_flag==Bad) 3	-0.022	0.0023	-0.026	-0.017	89.468	<.0001*
Prob(good_bad_flag==Bad) 2	Prob(good_bad_flag==Bad) 4	-0.024	0.0026	-0.029	-0.019	86.086	<.0001*
Prob(good_bad_flag==Bad) 3	Prob(good_bad_flag==Bad) 4	-0.003	0.0019	-0.006	0.0011	1.8745	0.1710

Test	ChiSquare	DF	Prob>ChiSq
All AUCs equal	1900.17	4	<.0001*

SMOTE paired with the Boosted Tree model provides the best performance.

After optimising the Boosted Tree algorithm, to reduce overfitting, it gives a misclassification rate of ~14%, accuracy of ~86%, and a low false positive rate of ~12%. The worst performing sampling methods are No Weighting, Tomek Majority NN Pairs and Tomek NN Pairs as they consistently gives high false positive rates across predictive models, which is undesirable for loan default prediction.

Conclusion and Future Work

- ✓ Best results was provided using SMOTE resampling technique paired with Boosted Tree Model. SMOTE produced the smallest false positive rates as compared to other sampling techniques.
- ✓ For loan default prediction, bank account type is the most important variable to include in the model (Savings Versus Current account type)
- ✓ Future work should include exploring other datasets like macro-economic indicators, optimising the Boosted Tree Algorithm, and researching the causes of variance in performance of resampling