Paper 3483-2015

Data sampling improvement by developing SMOTE technique in SAS

Lina Guzman, DIRECTV

ABSTRACT

A common problem when developing classification models is the imbalance of classes in the classification variable. This imbalance means that a class is represented by a large number of cases while the other is represented by very few. When this happens, the predictive power of the developed model could be biased given that classification methods tend to favor the majority class since they are designed to minimize the error on the total data set regardless of the proportions or balance of the classes.

Due to this problem, there are several techniques used to balance the distribution of the classification variable. One method is to reduce the size of the majority class (Undersampling), another is to increase the number of cases in the minority class (Oversampling) or to use a combination of the two. There is also a more complex technique called SMOTE (Synthetic Over-sampling Technique Minority) that consists of intelligently generating new synthetic registers of the minority class using a nearest neighbors approach.

In this paper we present the development in SAS of a combination of SMOTE and Undersampling techniques applied to a churn model. Then, we compare the predictive power of the model using this proposed balancing technique against other models developed with different data sampling techniques.

INTRODUCTION

The classification in two groups is one of the most common topics in information analysis involving data mining. In order to solve this problem, classification algorithms are developed to identify common patterns in a dataset, making it possible to get a classification model that can be generalized and applied to new datasets. However, this task is not simple and the problem is exacerbated by the imbalance of classes. This occurs when the number of instances in a class is much lower than the other class. A specific example of this behavior occurs in the telecommunications companies where the churn rate is lower than 3%.In these cases the classification models fail to capture the features of the minority class in the correct way and they end up classifying all customers in the same group.

To solve the problem of unbalanced classes, there are several strategies such as improving classification algorithms, or balancing classes in the sample training before developing the algorithm. The latter being the most used because it is independent of the applied method. Basically, balancing the classes consists of increasing the frequency of cases of the minority class or decreasing the majority class with the objective of obtaining the same number of instances in both groups.

The techniques of balance usually are:

Undersampling: Consists of deleting or randomly selecting cases from the majority class until the number of instances of the minority is balanced

Oversampling: Increases the number of instances in the class with a lower frequency in order to obtain a higher level of representation. Among the most known of these techniques are:

- Re-sampling: randomly duplicate instances of the minority class.
- SMOTE: Create synthetic or artificial records from instances of the group with lower frequency.

This paper presents application of SMOTE with different levels of increase, to a business case focused on management of costumers that want cancel the subscription to TV. The goal is to identify of early way customers who want make "churn" and apply preventive retention strategies. The paper is divides into three sections. First, there is an introduction to general concept necessaries for the developing the

application, second, a description of the data and the steps and finally the results are showing comparing the different levels of increase with undersampling in a test dataset for each model and the same dataset for all models.

GENERAL CONCEPTS

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE - SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is a proposed methodology which aims to reduce the effect of having few instances in the minority class. The strategy involves taking a subset of data from the minority class as an example and then intelligently creating new synthetic similar instances. These are then added to the original dataset, using the new dataset as a sample in the training process for the classifier model.

The SMOTE methodology creates the new synthetic dataset with these steps:

1. Randomly select a subset with proportion p from the minority class, where p<=100% is equal to the percentage of increase in the size of the minority class.

2. For each instance x selected in the subset at step 1, which is composed by n variables, select the k nearest neighbors inside the subset using Gower distance (which is discussed after the pseudocode).

3. Each variable of the synthetic instance is created progressively in this manner:

    3.1 Choose a random instance from the k neighbors

    3.2 Assign this value to the synthetic variable

    3.3 Repeat for each variable

In the case of continuous variables, step 3 is replaced by:

4. From the k nearest neighbors, select a random case y.

5. For each variable x subtract the value of the variable in y and multiply the result by a random number between 0 and 1. This is the new value that the synthetic variable will take.

The next pseudocode is used for continuous variables:

Algorithm *SMOTE*(T, N, k)

Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k

Output: (N/100)* T synthetic minority class samples

1. (**If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.**)
2. **if** N <100
3. **then** Randomize the T minority class samples
4. T = (N/100) *T
5. N = 100
6. **endif**
7. N = (*int*)(N/100)(**The amount of SMOTE is assumed to be in integral multiples of 100. **)
8. k = Number of nearest neighbors

9. *numattrs* = Number of attributes

10. *Sample*[ ][ ]: array for original minority class samples

11. *newindex*: keeps a count of number of synthetic samples generated, initialized to 0

12. *Synthetic*[ ][ ]: array for synthetic samples

(**Compute k nearest neighbors for each minority class sample only. **)

13. **for** *i* ←1 **to** *T*

14. Compute *k* nearest neighbors for *i*, and save the indices in the *nnarray*

15. Populate(*N*, *i*, *nnarray*)

16. **endfor**

Populate(*N, i, nnarray*) (**Function to generate the synthetic samples. **)

17. **while** *N* ≠0

18. Choose a random number between 1 and *k*, call it *nn*. This step chooses one of

the *k* nearest neighbors of *i*.

19. **for** *attr* ←1 **to** *numattrs*

20. Compute: *dif* = *Sample*[*nnarray*[*nn*]][*attr*] −*Sample*[*i*][*attr*]

21. Compute: *gap* = random number between 0 and 1

22. *Synthetic*[*newindex*][*attr*] = *Sample*[*i*][*attr*] + *gap *dif*

23. **endfor**

24. *newindex*++

25. *N* = *N* −1

26. **endwhile**

27. **return** (**End of Populate. **)

End of Pseudo-Code.

In the case of continuous variables, steps 18 onwards are replaced by:

18. **for** attr ←1 **to** numattrs

19. Choose a random number between 1 and k, call it nn. This step chooses one of

the k nearest neighbors of i.

20. Synthetic[newindex][attr] = Sample[nnarray[nn]][attr]

21. **endfor**

22. newindex++

23. N = N −1

24. **endwhile**

25. **return** (*End of Populate. *)

End of Pseudo-Code.

GOWER'S GENERALIZED COEFFICIENT OF DISSIMILARITY.

The Gower coefficient is a measure of similarity which can be defined between multivariate samples whenever the variables are a mixed type such as nominal, categorical or dichotomous, and continuous.

The Gower distance is defined as $d_{ij}^2 = 1 - s_{ij}$ where s_{ij} is the Gower coefficient of dissimilarity.

$$s_{ij} = \frac{\sum_{h=1}^{p1}(1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

p_1: Number of continual variables in the instance.

p_2: Number of binary variables,

p_3: Number of categorical variables (Not binary),

a: Number of matches (1, 1) in the binary variables,

d:Number of matches (0, 0) in the binary variables,

α:Number of matches in the categorical variables (Not binary),

G_h: Is the rank of the h[th] continuous variable.

COMPARISON MEASURES

In this study, were developed five models: four applying different levels of increase with SMOTE and one where only undersampling is applied. In order to assess the predictive performance of the classification models, various measures are calculated, these are explained below

RECEIVER OPERATING CHARACTERISTIC CURVE (ROC):

The ROC curve is one of the most used methods for evaluating the discrimination of a model. It is a graphic representation for all cutoffs of the Positive rate (sensitivity) plotted in function of the false positive (1-specificity).

The specificity and sensibility are calculated from the confusion matrix:

		Predicted value		
		Churners	**Not Churners**	**Total**
Real value	**Churners**	a	B	a+b
	Not Churners	c	D	c+d
	Total	a+c	b+d	n

Table 1: Confusion matrix

where:

$$Sensibility = \frac{a}{a + c}$$

$$specificity = \frac{d}{b + d}$$

KOLMOGOROV SMIRNOV (K-S):

The statistic (K-S) measures the absolute difference between distributions accumulated and determines the discrimination level of the model.
F_{Ch} *(s) and* F_{NCH} *(s)* are the cumulative distributions for each rank of definite probability of the churners and not churners respectively.

Where:

$$KS = max_s|F_{Ch}(s) - F_{NCh}(s)|$$

LIFT:

Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. for each rank *i* is calculated*:*

$$Lift = \frac{Bad\ rate_i}{Bad\ rate_{total}}$$

where:

$$Bad\ rate_i = \frac{frequency\ \ of\ bad_i}{total_i}$$

GINI:

It is a measure of overall quality of the model that is useful for comparing different models when the same dataset is used.

$$Gini = \sum_{i=1}^{n} d_i$$

Where:

$$d_i = \begin{cases} F_{Ch}(s)_i * F_{NCh}(s)_i\,, & i = 1 \\ [F_{Ch}(s)_i - F_{Ch}(s)_{i-1}] * [F_{NCh}(s)_i + F_{NCh}(s)_{i-1}], & i > 1 \end{cases}$$

DATA DESCRIPTION

The dataset is a subset of the population and has 364.865 client records, each record containing information about 178 categorical variables. These customers are classified as good or bad, where a bad client is one who has canceled their subscription, meaning they have "churned" and a good customer is one didn't. The composition of data is shown in the following table:

	Frequency	Percentage
Churners	4,813	1.32%
Non churners	360,052	98.68%
Total	364,865	100.00%

Table 2: Description Dataset

DESCRIPTION OF METHODOLOGIES

The oversampling technique SMOTE is applied to the minority class in different percentages of increment. At the same time, the majority class is reduced through undersampling until both groups have the same number of instances, as shown in Table 2.

Methodology	% Increment Churners	% diminution Non churners	Churners	Non Churners	Total records
Undersampling	0%	98.7%	4813	4813	9626
Smote + Undersampling	25%	98.3%	6017	6017	12034
	50%	98.0%	7220	7220	14440
	75%	97.7%	8423	8423	16846
	100%	97.3%	9626	9626	19252

Table 3: Increment and disminution of data.

TEST PARAMETERS

- For each dataset three subsets are created: training, validation and testing. Sixty percent for training and 20% for validation and testing.
- Number of neighbors: 5
- Distance function: Gower
- The majority class is reduced to frequencies of 50% in each group.
- Generate a logistic regression model for each balanced dataset.
- Calculate the comparison measures for each model.

RESULT

Each model was tested on two different data sets, in the validation sample and out-of-sample data.

TEST IN VALIDATION SAMPLE

The comparison of the ROC curves for each model shows that SMOTE in the minority class combined with undersampling in the majority works better than only undersampling in the validation sample.

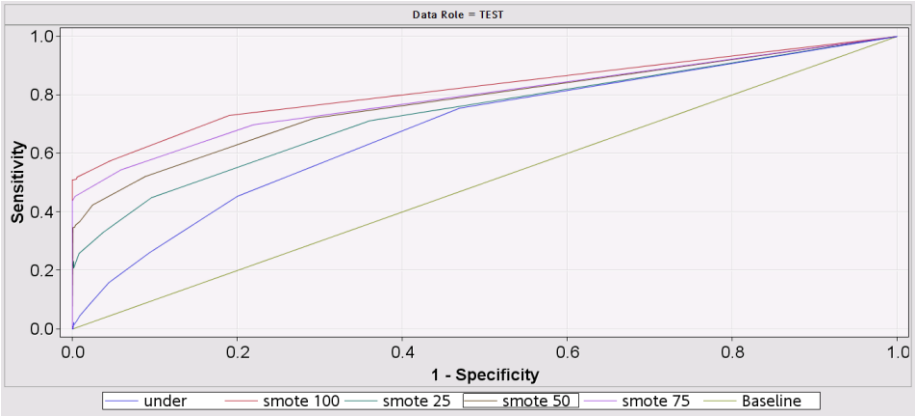

Figure 1: Curve ROC - Dataset Test

% Smote	ROC	GINI	KS
100%	84.70%	69.40%	53.70%
75%	81.90%	63.80%	48.40%
50%	80.30%	60.60%	43.20%
25%	76.80%	53.60%	35.20%
Undersampling	69.90%	39.80%	28.50%

Table 4:Measures of comparison Data set test

The models generated with a percentage increase of the minority class using SMOTE have a higher level of discrimination against the models generated using undersampling in the majority class. (Table 4)

TEST IN OUT-OF-SAMPLE

Tests with data out-of-sample were carried out by applying the models generated at the same dataset, in Figure 2. At first glance the differences in the ROC curves seem to be very small.

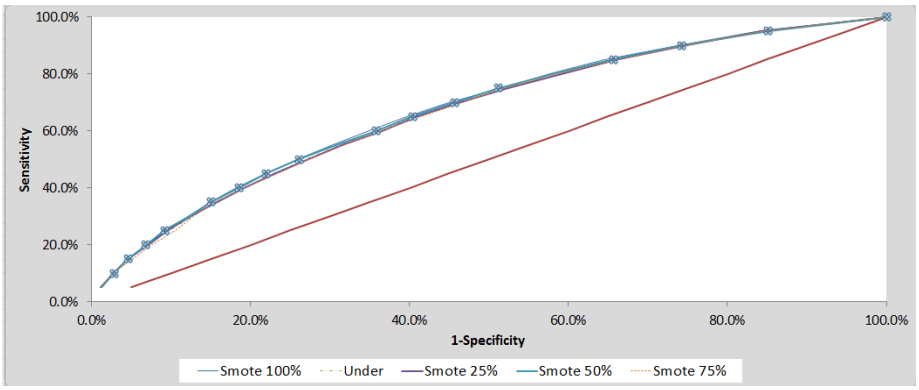

Figure 2: ROC Curve, out of sample

Table 4 shows the significant improvement of the discrimination (KS) of the models generated with SMOTE and undersampling and those which only applied undersampling. Additionally, a higher number of bad customers is captured with the 20% of the population the model created with SMOTE, resulting in increased effectiveness of the model in terms of customer management.

% Smote	KS	Cumulative Response at 20% population	Lift
100%	25.7%	42.3%	3.18
75%	23.9%	41.4%	3.11
50%	24.9%	42.0%	3.02
25%	24.3%	41.0%	3.14
Undersampling	24.6%	40.9%	2.98

Table 5: Measures of comparison, out of sample

The Lift Curve Figure 3. helps analyze the amount of real churners who are discriminated against in each quintile. Table 5 shows the lift for each rank in the 5 models. In the first four quintiles that have the customers with are more likely to churn, the lift for the models with SMOTE is highest. On the other hand, in the quintiles 17 to 20, the lift is lower, showing a higher sensibility in the quintiles of lower risk.

Rank	Under Sampling	Smote 25%	Smote 50%	Smote 75%	Smote 100%
1	2.985	3.138	3.020	3.105	3.179
2	2.173	1.882	2.149	2.000	2.083
3	1.746	1.725	1.710	1.630	1.846
4	1.285	1.447	1.536	1.516	1.355
5	1.380	1.343	1.355	1.333	1.193
6	1.240	1.178	1.098	1.035	1.362
7	1.044	1.079	1.058	1.072	1.053
8	0.952	0.891	0.930	0.970	0.940
9	0.921	1.015	1.018	0.934	0.952
10	0.887	0.845	0.910	0.908	0.885
11	0.830	0.838	0.832	0.791	0.790
12	0.720	0.786	0.683	0.829	0.725
13	0.698	0.676	0.697	0.730	0.618
14	0.611	0.661	0.551	0.546	0.605
15	0.576	0.570	0.620	0.453	0.551
16	0.533	0.495	0.476	0.607	0.465
17	0.507	0.522	0.445	0.484	0.480
18	0.336	0.342	0.354	0.475	0.379
19	0.345	0.337	0.310	0.330	0.308
20	0.231	0.227	0.258	0.206	0.233

Table 6: Lift in each quintile

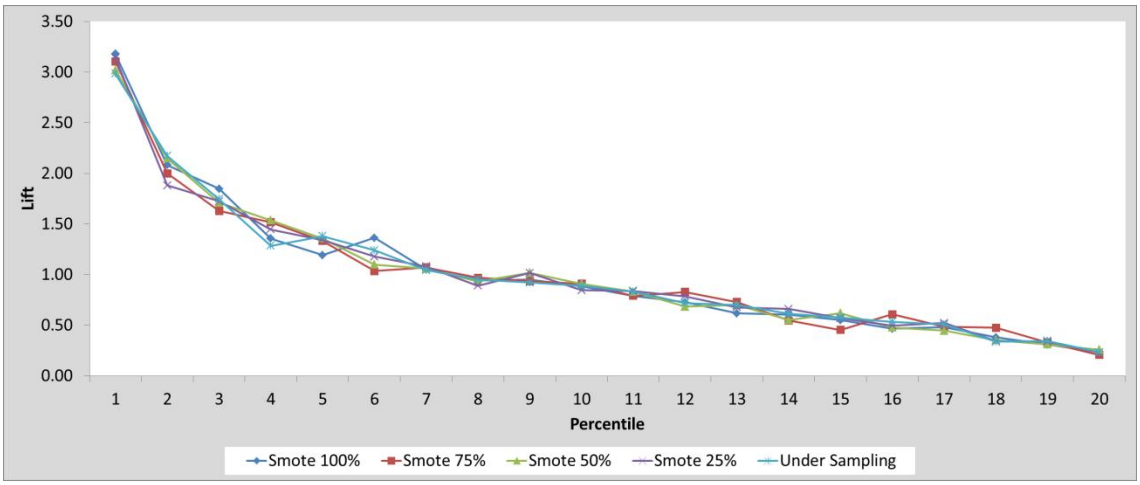

Figure 3: Lift Curve

CONCLUSION

The exercise results shown that applying the algorithm SMOTE in the minority class combined with undersampling in the majority, is possible improve the discrimination level and increase the specificity and sensibility of the model, in practical terms, helps to identify with major exactly the customers with higher propensity to make churn, additionally, with the applied of SMOTE, the percentage of diminution for the majority class is lower, this decreases the lost of information because is necessary delete less cases to achieve the balance.

Normally, this methodology used the Euclidian distance for find the near neighbors, however this measure of similarity, is only valid for continuous or numeric variables, and in a business context the data usually are mixed type. This problem took us to search for a measure of similarity that adapts to for different types of variables, for this reason the Gower coefficient is the measure most indicated.

REFERENCES

NiteshV.Chawla, Nathalie Japkowicz and AleksanderKolcz.: Editorial: Special Issue on Learning from Imbalanced Data Sets.

Chawla Nitesh, Bowyer Kevin, Hall Lawrence and Kegelmeyer Philip SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence and Research 2002 pp. 321-357

EnislayRamentol·Yailé Caballero · Rafael Bello ·Francisco Herrera SMOTE-RSB∗:a hybrid preprocessing approach basedon oversampling and undersampling for high imbalanceddata-sets using SMOTE and rough sets theory. 2011

SMOTE-RSB: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory

Sandrine Pavoine, Jeanne Vallet, Anne-Beátrice Dufour, Sophie Gachet and Hervé Daniel: On the challenge of treating various types of variables: application for improving the measurement of functional diversity

Teemu Mutanen, Customer churn analysis – a case study

RECOMMENDED READING

- *Base SAS® Procedures Guide*

- *SAS® For Dummies®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name	Lina María Guzmán Cartagena
Enterprise	DIRECTV Colombia
Phone	(+57) 3204016785
E-mail	linguz@directvla.com.co/lglinaguzman@gmail.com