SAS® GLOBAL FORUM 2020 #SASGF

Paper 5021-2020

Modelling imbalanced classes

Humphrey Brydon, Rénette Blignaut, Retha Luus, Isabella M Venter, University of the Western Cape; Desireé J Cranfield, Swansea University

ABSTRACT

In this study separate sampling was applied to various modelling procedures to assist in the identification of the most important variables describing smartphone users who are security compliant. Initial analysis of the data found that only 7% of smartphone users reported applying security measures to protect their phones and/or their personal information stored on their devices.

Due to the class imbalance in the target variable, predictive modelling procedures failed to produce accurate models. Separate sampling proportions were introduced to establish if classification accuracy could be improved. This study tested target class over-sampling ratios of 20%, 30%, 40% and 50% and compared the results of the models fitted on these data sets to those fitted on the original data where no separate sampling was applied.

Models fitted included: decision trees, 5-fold cross-validated decision trees, logistic regression, neural networks and gradient boosted decision trees. The results showed that the logistic regression and neural network models produced unstable models regardless of the target class ratios. More stable models were however reported for the decision trees, 5-fold cross-validated decision trees and gradient boosted decision trees.

Variables found to influence mobile security compliance included age, gender and various security/privacy related behaviors.

INTRODUCTION

According to a study conducted by the Pew Research Center, more than 5 billion people worldwide own a mobile device, with more than half of these mobile devices being smartphones (Silver and Taylor, 2019). In their study, conducted in 2018, it was also found that smartphone users in advanced economies versus those in emerging economies, were more likely to access the internet and social media. It is further reported that the number of smartphone users in emerging economies is increasing and this could therefore lead to more users globally accessing the internet and social media via their smartphones.

According to a study conducted by Strategy Analytics and as reported by Dechen et al. (2020), the mobile workforce is expected to increase to 1,87 billion people in 2022 (*from 1.45 billion in 2016*) or 42,5% (*from 38.8% in 2016*) of the global workforce. However, with the increase in mobile technology as well as more and more companies shifting to cloud-based systems, this increase in the mobile workforce is however, not surprising.

With the forecasted increase in not only the mobile workforce but also the number of people who use smartphones, the need for a proportional increase in smartphone or mobile device security technology is also needed. As reported by Becher et al. (2011), "real attacks" on smartphones started making headlines in 2010, although this is not to say that threats to mobile devices were not around prior to 2010.

In their study they further go on to identify four types of threats to mobile devices:

- “Hardware-centric attacks”;
- “Device-independent attacks”;
- “Software-centric attacks”; and
- “User layer attacks”.

In the study conducted by Dechen et al. (2020) they state that of the Chief Information Officers (CIO) surveyed, 81% reported that all WiFi related security incidents occurred outside of their company inside cafés and coffee shops (i.e. the mobile workforce accessing/using open public networks).

Therefore, with a possible mobile workforce of 1,87 billion in 2022 the question that should be asked is: Are smartphone users security compliant with respect to their smartphones?

OBJECTIVE

The data collected for this study combined four study cohorts where the same study design and questionnaire were used. After initial analysis of the data (*with n = 448*), it was found that only 7% of smartphone users reported applying security measures to their smartphones to secure not only their smartphones but also their personal information. Due to the imbalance in the target class, the predictive models initially used, failed to produce stable and accurate models.

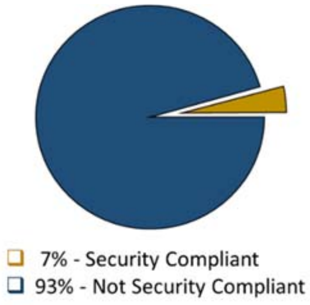

Figure 1. Ratio of security compliant users

In this study the effect of applying separate sampling techniques to the original data, is considered. The ultimate goal of doing so is to produce more stable and accurate predictive models that could assist in the identification of smartphone users who were security compliant.

SEPARATE SAMPLING

Since the original data set contained binary target proportions of 7% (*i.e. security compliant users*) and 93%, separate sampling was needed as the initial predictive models constructed were unstable at this proportion. An oversampling technique was used to produce various data sets with a 20%, 30%, 40% and 50% proportion in the target class of interest so that more stable and accurate models could hopefully be constructed.

The proc surveyselect procedure in SAS® was used to create the 4 new data sets. An example of the code used to create the 20% oversampled data set is given below (*the values 358 and 90 refer to the total number of observations in each target classifier respectively*):

```
proc surveyselect data = “init_data” out = “new_data” method=urs
            sampsize=(358 90);
     strata “target”;
run;
```

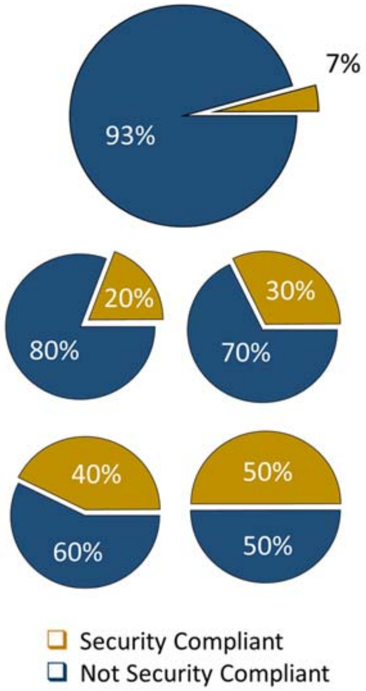

Figure 2. Ratio of security compliant users in new data sets

All models constructed in this study, regardless of the separate sampling ratio used, were tested on training and validation data using an 80:20 split respectively. For the data sets that were constructed by oversampling the target classifier, a prior probability was defined in SAS® Enterprise Miner in order to adjust for the population target proportions.

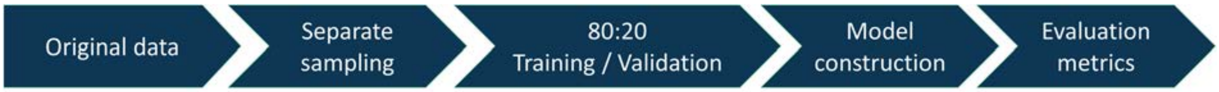

Figure 3. Approach taken in the analysis

MODELLING PROCEDURES

Five different modelling procedures were used in this study to predict the target classifier, “security compliant”. These models included, a decision tree, 5-fold cross-validated decision tree, a logistic regression model, a neural network as well as a Stochastic Gradient Boosted Decision Tree (SGBT).

The basic decision properties were set in order to perform binary splits, with a maximum tree depth of six. The splitting criterion in the decision tree was evaluated using the Gini purity measure. The cross-validated decision tree also performed binary splits at each parent node and contained the same depth and splitting criterion of the basic decision tree except that cross-validation was performed for this decision tree. A 5-fold approach was taken for the cross-validated decision trees.

The logistic regression model was constructed using a stepwise variable selection method. No interaction or polynomial terms were used in the construction of the logistic regression model.

The neural network was constructed using the more commonly used, Multilayer Perceptron architecture. The architecture contained a single hidden layer with three hidden units. The hyperbolic tangent function was used as the activation function in this modelling procedure. For the SGBT model, a total of 200 iterations were used in the construction of this decision tree, applying a 0.1 learning rate and 60% sampling rate for each iteration.

In total, 25 models were constructed (*i.e. models constructed on both the original and 4 newly created data sets*). The results for the models are discussed in the following section.

RESULTS

Of the five models constructed on the original data set, none produced a final model. Although fit statistics were produced here, these were found to be immaterial as the final models contained no variables and therefore these results are excluded from the final analysis and interpretation.

All models were compared using 4 fit statistics, namely; Misclassification Rate, ROC Index, Gini Coefficient and Kolmogorov Smirnov Statistic. These are given in Figure 3.

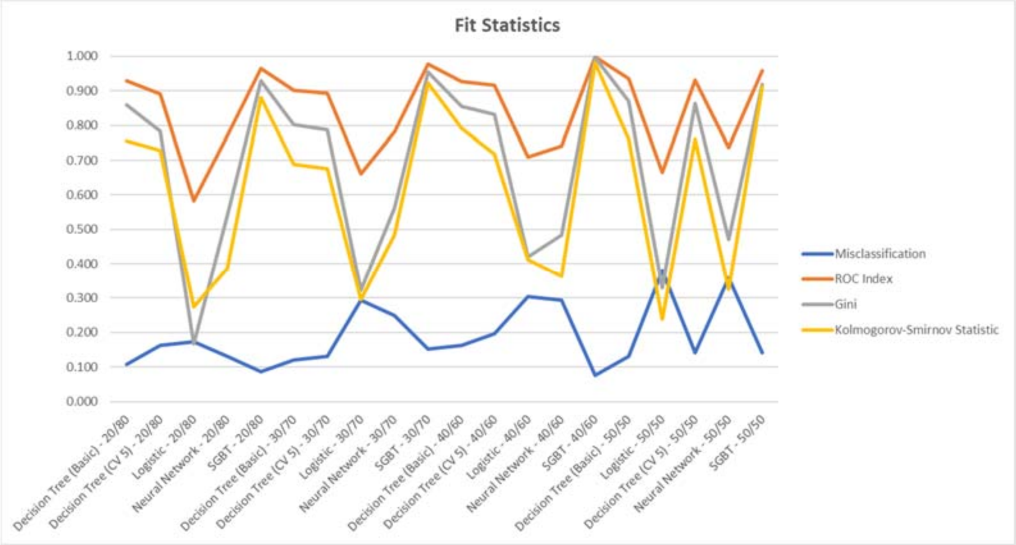

Figure 3. Model fit statistics

From Figure 3 it can be seen that the misclassification rates for all models were quite low. These results were however found to be unreliable due to the addition of prior probabilities in the modelling procedure and hence were not used in the final selection of a model.

From Figure 3 it can also be seen that for each oversampling ratio's set of models, the logistic regression model in that set produced the worst results, indicating that the logistic regression model did not perform as well as other models under oversampling. It was also found that the neural network tended to drop off in accuracy for the 40% and 50% oversampled data sets as compared to other oversampling ratios used in this study.

Although all the decision tree based modelling procedures performed well and with some stability across all the data sets, it was found that the SGBT models produced the most stable and accurate fit statistics across all data sets. A summary of these results are given in Figure 4.

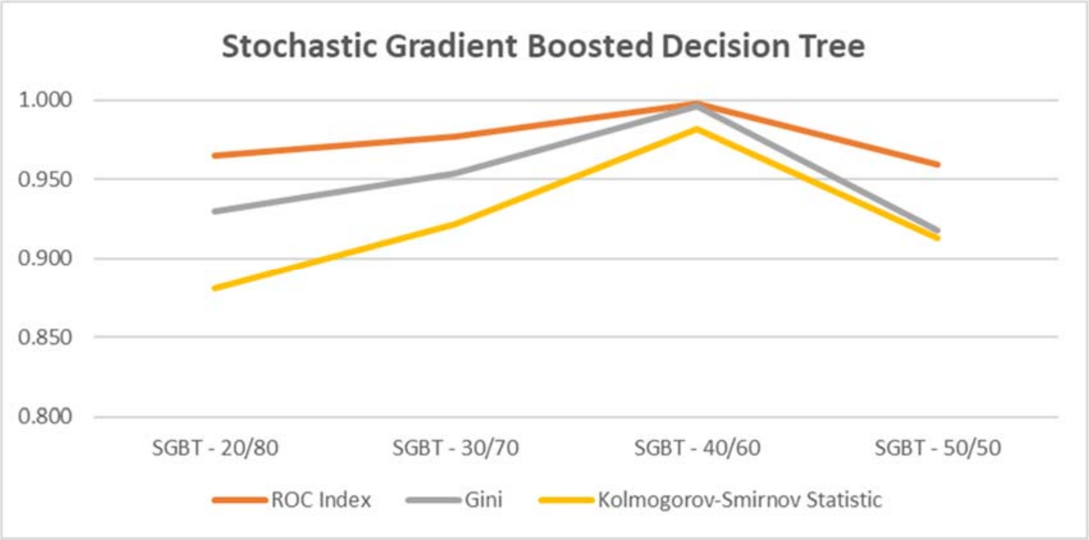

Figure 4. SGBT results

As can be seen from results summarized in Figure 4, the SGBT model constructed on the 40% oversampled data produced the best fit statistic results overall. This model produced a final model with seven variables, from the initial 50 predictor variables included in the modelling procedure.

Therefore, the variables that were found to be significant in determining whether smartphone users were security compliant were:

- The university that the user attended;
- Whether the user felt social media provided enough information about their smartphone privacy and security issues;
- Where/who the user gets advice from regarding privacy and/or security issues;
- Whether the user felt that social media posts could be tainted by other social media posts;
- The duration of ownership of a smartphone; and
- Whether the user felt that instant messaging services had security/privacy implications.

CONCLUSION

It was found in this study that a small proportion in the target classifier resulted in predictive models failing to produce reliable results when constructed on data with a small representation in the target classifier. In all of these cases, the final model also contained no variables. Although some of the models constructed could be considered as advanced statistical modelling procedures, they could still not overcome the sparsity of the target classifier in the data.

Oversampling ratios of 20%, 30%, 40% and 50% on the target classifier were included in the separate sampling procedure carried out on the original data set. Prior probabilities were also defined in the modelling procedure so as to take into consideration the sparsity of the target classifier in the original data set.

All models constructed on the oversampled data showed an increase in model accuracy with the SGBT models showing the most improved, stable and accurate results across all data sets. The SGBT model constructed on the 40% oversampled data set however produced the best results of all models across all data sets.

The results showed that the seven variables that could be significant in determining whether a smartphone user was security compliant were related to the university they attended, the user’s preferences regarding social media and instant messaging privacy/security issues, where the user gets advice from and also the length of time that the user has had a smartphone.

REFERENCES

Silver, L. & Taylor, K. (2019). *Pew Research Center - Smartphone Ownership is Growing Rapidly Around the World, but Not Always Equally,* Pew Research Center.

Luk, G. & Brown, A. (2016). *Strategy Analytics.* [Online]
Available at: https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/2016/11/09/the-global-mobile-workforce-is-set-to-increase-to-1.87-billion-people-in-2022-accounting-for-42.5-of-the-global-workforce
[Accessed 14 February 2020].

Dechen, T. et al. (2020). A preliminary study of risk assessment of mobile workers for improvement of work-life balance. *Bulletin of Networking, Computing, Systems, and Software,* 9(1), pp. 43 - 45.

Becher, M. et al. (May 2011). *Mobile Security Catching Up? Revealing the Nuts and Bolts of the Security of Mobile Devices.* Berkeley, CA, IEEE: Symposium on Security and Privacy.