Taxpayer Behavior Prediction in SMS Campaigns

Guiming Cao, Alan Downes, Shuraia Khan, Wendy Wong, Guandong Xu
Advanced Analytics Institute
University of Technology Sydney, Australia
{Guiming.Cao, Shuraia.Khan, Wendy.Wong-1}@student.uts.edu.au
{Alan.Downes, Guandong.Xu}@uts.edu.au

***Abstract*—This paper considers a prediction study of a group of small businesses which have a higher risk of non-compliance with taxation obligations. These businesses have been selected for a pre-emptive SMS reminder campaign and prediction models are used to predict the probability of on-time payment. Through experiments on a real world taxation debt dataset, it is found that the XGBoost algorithm significantly outperforms random forest, decision tree and logistic regression algorithms. The variables showing the largest explanatory power are related to debt amount. Second and subsequent SMS messages make a negligible contribution to the probability of payment. The XGBoost explainer is also used to delve further into the inner workings of the algorithm.**

***Index Terms*—Taxpayer Behavior, Debt Collection, SMS, Prediction, XGBoost**

I. Introduction

Over the last twenty five years mobile phones have become ubiquitous in the community. The use of SMS messaging enables companies to have a much closer and more immediate relationship with their customers compared to the more traditional approach of sending letters. Recent research has focused on nudging consumers to engage in appropriate behavior [1]. SMS messaging can assist in this, through issuing appointment [2] and debt reminders [3].

SMS can also be used by governments for specific purposes, such as debt reminders. For example, the Australian Taxation Office has engaged in a trial study examining the effectiveness of sending SMS messages before the due date to those with a history of poor compliance. The business problem is how to choose the right population to make such SMS campaigns cost-effective. Such nudges are only likely to be effective for those with a reasonable likelihood of on-time payment. Therefore, finding the most powerful prediction model and examining how well it works in this context is the key research question addressed in this study. Over past decades, prediction models which make use of classfication algorithms such as decision trees have become very popular in the machine learning community. Starting with the seminal paper of Breiman in 2001 [4], so-called "random forests", where a collection of decision trees vote on the best solution, have been very successful in prediction modeling. Better-performing variants of these models have been developed in recent years, the XGBoost model developed by Chen and Guestrin [5] being one of the most successful. This paper augments that study by using an XGBoost model in an attempt to predict which of these higher-risk taxpayers will pay their obligations on time. Results are compared with simpler logistic regression, single decision tree and random forest approaches. The question of which variables are the most useful predictors is also examined.

This study was funded by the Australian Taxation Office and an ARC Linkage Industry grant.

The SMS campaign used in this study was conducted within the Australian Taxation Office (ATO). Taxpayers were selected for inclusion in the SMS trial based on whether the probability of paying on time was less than a 49% threshold. These probabilities were calculated using a in-house Payment On-Time model using logistic regression model to analyse Activity Statement compliance. Pre-due date SMS messages were sent five days before the lodgment date. If the taxpayer did not pay on time, then they were also sent a post-due date SMS within seven days of the due date. Only the pre-due date messages were considered in this study.

The Organisation for Economic Co-operation and Development (OECD) has examined the criteria used to select clients by the ATO [6] [7]. The ATO has used analytics within the debt collection process and to segment taxpayers based on their engagement with the ATO [7]. The results are used as input to policy reforms and to identify emerging risks.

This paper makes the following contributions:

- Addresses the effectiveness of an SMS compaign in a taxation context using prediction modelling.
- Tests the effectiveness of various prediction algorithms in predicting taxpayer behavior. Amongst these algorithms, the XGBoost algorithm [5] performs the best and shows greater predictive power than the alternatives considered.
- Showcases the XGBoost explainer [8] [9], which has recently been developed to clarify how the XGBoost algorithm operates internally.

II. Related Work

Tax-related analytics has increased in importance in recent years. The Tax and Customs Administration of the Netherlands have adopted a "Chain of collection process" to select the client based on a series of chains of action outlined

in OECD (2014) [6].

The Belgium Tax Administration [6] has built models to predict the likelihood of insolvency. Analytic approaches included risk analysis and risk ranking for predicting insolvency risk and payment behavior.

As outlined in [6], the Australian Tax Office (ATO) has built risk models to classify taxpayers and work out the most appropriate debt collection method to apply. Inputs include lodgment and payment history. The final risk score for a taxpayer is built up by modeling the propensity to pay (likelihood of clearing all debts) and capacity to pay (considers the likelihood of insolvency). The ATO has also developed assessment tools to prevent debts arising in the first place [6].

When sending SMS text messages, personalizing the message is important in getting a favorable response. Humbani [10] concluded in the context of SMS advertising that consumers respond more positively to informative and personalized messages. The London Cabinet Office [11] found 10% more people paid outstanding court fees after receiving a personalized text message.

The XGBoost algorithm has been used in applications as varied as building an indoor positioning system [12] and predicting customer churn rates for an online music service [13].

III. Dataset

The dataset targeted taxpayer behavior with a focus on client interactions for up to two years prior to receiving a pre-due date SMS, subsequent interactions, compliance and monitoring of behavior post SMS delivery. The variables used were intended to give an overall picture of the taxpayer, their interactions with the ATO and subsequent behavior as a result of receiving the SMS.

A. Target Population

Small business taxpayers were selected who had received a pre-due date SMS. Behavior was analyzed for a period of 2 years prior to the first SMS message (2015 to 2017 financial years).

B. Variable Selection

The variables selected for modeling are sourced from four datasets:

1) Demographics: Applied to understand taxpayers better who have received an SMS. Trends and commonalities might be identified that can feed into future engagement strategies in targeting this population for preventative or predictive strategies.

2) Debt and Payment: These variables help determine the effect on the taxpayer receiving the SMS and if payment is made. It also helps form a picture of the taxpayer by assessing their debt profile. The debt profile is used in conjunction with lodgment information to develop an overall compliance profile of taxpayers who receive an SMS.

3) Interactions: Includes data on SMS messages sent and other interactions that the taxpayer has had with the Australian Taxation Office. This helps to highlight differences in engagement with the ATO pre and post SMS.

4) Lodgment: This allows assessment of whether on-time lodgment improves after sending an SMS message.

C. Data preparation

The dataset comprised around 460K observations. 80K observations were selected from these and used to train, validate and test the model. The train / validation / test split was 60/20/20. Five days before the lodgment due date, the taxpayer received an SMS message. The target indicator variable is then built, showing whether the debt has been paid off thirty days after the lodgment due date (value 1) or not paid off (value 0).

IV. Methods

A series of prediction models were built to evaluate the outcomes of sending preventative SMS messages and to identify if SMS messages were driving longitudinal behavioral change or if they lose their effectiveness. In this study, we explore logistic regression, random forests, decision trees and XGBoost.

Logistic Regression. The baseline algorithm for the classification research problem was logistic regression.

Decision Tree The most basic tree-based approach to prediction modeling is the decision tree. The aim of this method is to build up a tree of classification steps depending on the values (or "split points") of selected predictor variables. Variable and split point selection is performed automatically.

Random Forest. The random forest algorithm was introduced by Breiman [4]. The main idea is to use a collection of trees rather than a single decision tree in order to improve prediction accuracy. Each individual tree in the forest has a vote on what the final prediction for an observation will be.

XGBoost. Boosting also uses a collection of decision trees, but iterates towards a solution using an optimization algorithm known as gradient descent [14]. XGBoost [5] takes this a step further, producing a scalable boosting system with specific allowance for sparse data. It also considers second order derivatives in the optimization algorithm [15, p.68]. This system has been highly successful in machine learning

competitions [5, p.1].

More technically, boosting aims to fit a succession of functions $\phi_m(x)$ to build an overall predictive function $f(x)$ [15, p.39]:

$$f(x) = \theta_0 + \sum_{m=1}^{M} \theta_m \phi_m(x)$$

where θ_m, $m = 0, \ldots, M$ is a set of weights.

The standard boosting algorithm chooses each $\phi_m(x)$ as follows [15, p.40]:

$$\hat{\phi}_m(x) = \underset{\phi(x),\beta}{\arg\min} \sum_{i=1}^{n} \left[(-\hat{g}_m(x_i)) - \beta\phi(x_i)\right]^2$$

where $\hat{g}_m(x_i)$ is the gradient of the expected loss function at the ith data point x_i [15, p.36] and β is a parameter to be optimised.

XGBoost modifies this by also taking into account the second derivative $h_m(x)$ of the expected loss function [15, p.42]:

$$\hat{\phi}_m(x) = \underset{\phi(x)}{\arg\min} \sum_{i=1}^{n} \frac{1}{2}\hat{h}_m(x_i) \left[\left(-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right) - \phi(x_i)\right]^2$$

One issue historically with the XGBoost algorithm is that it has been difficult to explain how the predictions are determined. It has often been treated as more of a black-box algorithm. This issue has recently been resolved with the development of the XGBoost explainer [8] [9]. This algorithm aggregates across all of the trees to produce a breakdown of individual variable contributions to a prediction, similar to that available with a single decision tree.

V. Results

This paper applies four machine learning models, namely logistic regression, decision tree, random forest and XGBoost in order to predict taxpayer compliance behavior. The XGBoost model utilized an R package [16], with the other models being run in SAS Enterprise Miner.

The misclassification rates for the various approaches, as shown in Fig. 1, are: logistic regression: 0.34, decision tree: 0.28, random forest: 0.28 and XGBoost 0.25.

The AUC measures, as shown in Fig. 2, are: logistic regression: 0.51, decision tree: 0.69, random forest: 0.69 and XGBoost 0.81.

All of the tree-based models perform better than logistic regression in terms of both misclassification rate and AUC. XGBoost performed the best on both of these measures. The receiver operating characteristic curve (ROC) and AUC for XGBoost are shown below in Fig. 3.

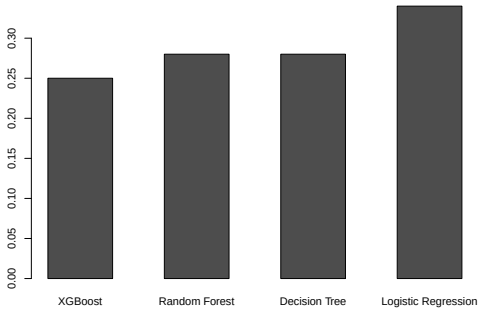

Fig. 1. Misclassification - all models.

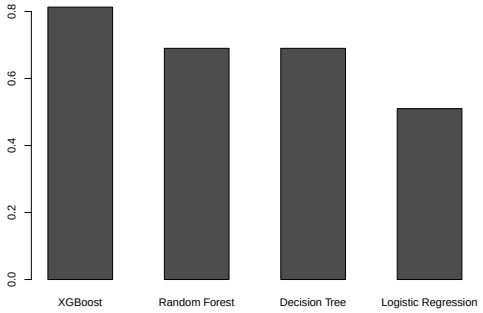

Fig. 2. AUC - all models.

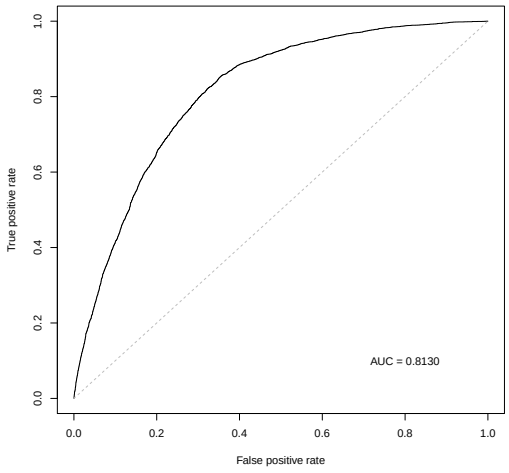

Fig. 3. XGBoost ROC and AUC.

The importance graph in Fig. 4 shows that the three debt-related variables make the largest contribution to the model's overall explanatory power. These variables differ in how credits and payments are taken into account in the debt value. Whether or not the business has employees has a fairly small impact.

Other variables considered include whether the business is registered for GST, whether the business is a sole trader, partnership or public company and whether the business uses a tax agent. Categorical variables were also included for how many SMS messages were sent (up to five) and the classification of the business according to the ATO's internal risk system. The contribution of all of these additional variables is negligible. In particular, second and subsequent SMS messages don't change the probability of on-time payment.

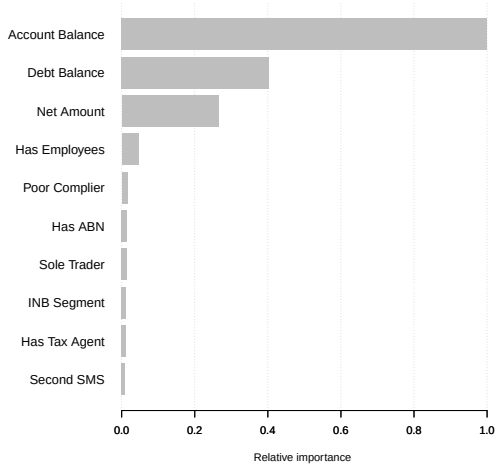

Fig. 4. Variable importance.

Using the XGBoost explainer, Fig. 5 shows that those companies who have employees (colored black) are more likely to pay their debt on time than those who don't have employees (colored gray). This makes sense as businesses which have employees are likely to be larger and more stable, having the systems in place to ensure their obligations are met on time.

The XGBoost explainer can also be used to show how variables are contributing to an individual prediction, as shown in the waterfall chart in Fig. 6 for a hypothetical small business. Starting at the 0.5 position, each successive variable makes a positive or negative contribution to the probability. The bar positions are cumulative - the starting position for a particular bar is the same as the ending position of the previous bar. The contributions made by each variable get smaller and smaller until the final position (the prediction) is reached. The final bar shows the overall probability of

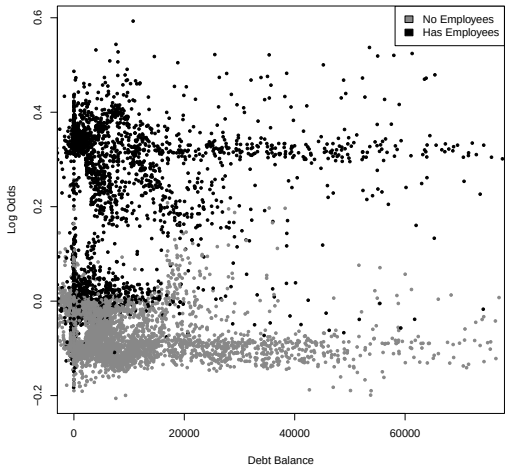

Fig. 5. Payment likelihood (log odds) vs debt balance.

payment for the business. The additive modelling itself is performed using log odds ratios and these are converted to probabilities using the sigmoid function.

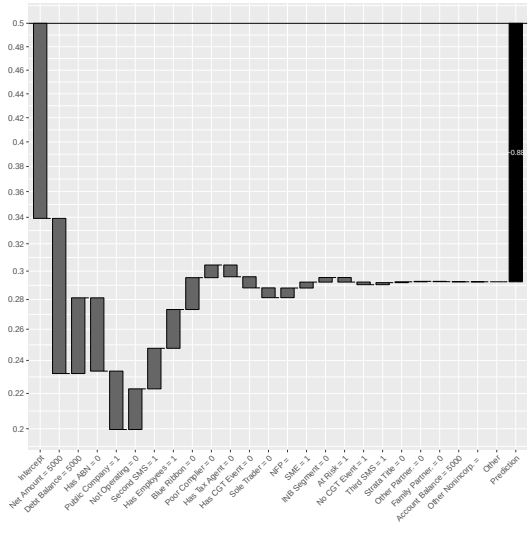

Fig. 6. Waterfall chart - hypothetical taxpayer.

VI. Conclusions

In this study, we have investigated the use of prediction models for an SMS campaign. For those higher risk small business taxpayers which are the subject of this study, the most important predictor of whether or not payment will be made is the amount of the debt. Subsequent SMS messages after the first one have minimal influence on the probability of payment. This study also highlights the effectiveness of the XGBoost algorithm for these sorts of investigations. It achieved an absolute improvement in AUC of 30% above the baseline logistic regression model. It also shows an improvement of

12% above the decision tree and random forest approaches. Finally, this paper shows how the XGBoost explainer can be used to see which variables are contributing the most to the probability of payment for a single business taxpayer.

REFERENCES

[1] R. H. Thaler and C. R. Sunstein, *Nudge: Improving Decisions about Health, Wealth and Happiness*. Yale University Press, 2008.

[2] E. Koshy, J. Car, and A. Majeed, "Effectiveness of Mobile-Phone Short Message Service (SMS) Reminders for Opthalmology Appointments: Observational Study," *BMC Opthalmology*, vol. 8, no. 9, 2008.

[3] P. Pekonen, "Are Text Message Reminders Effective in Debt Collection? Randomized Controlled Trial in Debt Collection in Finland," Masters Thesis, Department of Finance, Aalto University, 2014.

[4] L.Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD*, 2016.

[6] OECD, *Working Smarter in Tax Debt Management*. OECD Publishing, 2014.

[7] OECD, *Behavioural Insights and Public Policy: Lesson from Around the World*. Paris: OECD Publishing, 2017.

[8] D. Foster, "New R Package that makes XGBoost Interpretable," 2017. [Online]. Available: "https://medium.com/applied-data-science/new-r-package-the-xgboost-explainer-51dd7d1aa211"

[9] R Package, "XGBoost Explainer." [Online]. Available: https://github.com/AppliedDataSciencePartners/xgboostExplainer

[10] M. Humbani, T. Kotzé, and Y. Jordaan, "Predictors of Consumer Attitudes Towards SMS Advertising," *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, vol. 24, no. 2, pp. 2–19, 2015.

[11] London Cabinet Office, "Applying Behavioural Insights to Reduce Fraud, Error and Debt," Behavioural Insights Team, London, UK, Tech. Rep., 2012.

[12] M. Luckner, B. Topolski, and M. Mazurek, "Application of XGBoost Algorithm in Fingerprinting Localisation Task," *Lecture Notes in Computer Science*, vol. 10244, pp. 661–671, 2017.

[13] B. Gregory, "Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data," *ArXiv e-prints https://www.arxiv.org/pdf/1802.03396.pdf*, 2018. [Online]. Available: "http://www.arxiv.org"

[14] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[15] D. Nielsen, "Tree Boosting With XGBoost," Masters Thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology, 2016.

[16] R Package, "xgboost." [Online]. Available: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf