HIGHLIGHTS FROM AN INDEPENDENT STUDY OF CREDOLAB'S SCORING ALGORITHM

MACHINE LEARNING-DRIVEN CREDIT RISK MODELLING USING SMARTPHONE METADATA

Xiaofei (Susan) Wang, PhD
Former Lecturer and Research Scholar
Department of Statistics & Data Science
Yale University

credolab

WHAT WE'LL TALK ABOUT

Introduction

Abstract

The Data

The Scoring Algorithm - Dataset Organization

The Scoring Algorithm - Feature Selection

The Outcome

The Experiment

Conclusion

INTRODUCTION

The 21st century is marked by the democratisation of the internet for mass consumption. In this day and age, personal cell phone devices are no longer luxury items only afforded by the rich - indeed, as of 2017, an estimated 66% of the world population owned a cell phone and about 57% of cell phone users owned smartphones. These quantities are projected to grow to 71% and 77%, respectively, by 2025 (Sivakumaran & Iacopino, 2018).

As smartphones facilitate day-to-day interactions such as calling, texting, emailing, and more, they have become more than just windows to the outside world; they are also an electronic diary of individual phone users. A recent working paper by the Federal Deposit Insurance Corporation's Center for Financial Research found that even a minimal set of digital footprint data including features like device operating system and email host could yield predictive models for default (Berg, Burg, et al., 2018).

71%

cell phone owners in the world by 2025 from 66% in 2017*

77%

smartphone users in the world by 2025 from 57% in 2017*

*The Mobile Economy 2018. GSMA Intelligence

ABSTRACT

Credolab is at the forefront of innovative risk management practices that engage with novel credit risk modelling approaches availed by the surge in cell phone use. Core to credolab's business is its modelling pipeline. Taking the smartphone as input, the data processing pipeline consists of a series of automated steps, rooted in machine learning techniques, that ultimately outputs a predictive model for credit default. To protect confidentiality and to ensure against bias towards individual loan customers, only non-identifying metadata is used.

This paper reports the findings of an independent review of credolab's scoring model. The independent review considered a vast array of alternative approaches for the various different steps of the pipeline and found favourable results, including when applied to real data. We first explore the data sets that credolab consumes, how it translates it into scores, and the outcome it serves. In the latter part of the paper, we take a look at how credolab's algorithm fared when compared to that of other major players with similar scoring models.

THE DATA

The user’s smartphone data form the units of data that goes into credolab's modelling pipeline. Observed input variables include aggregated summary statistics such as the amount of time the user spent on phone calls over the past month. Some of these summary statistics are on a continuous scale (such as minutes called); others are on a discrete scale (such as the frequency of calls) and still, others are binary (such as whether particular phone applications are installed). The number of observed variables varies based on availability but could easily run in the thousands.

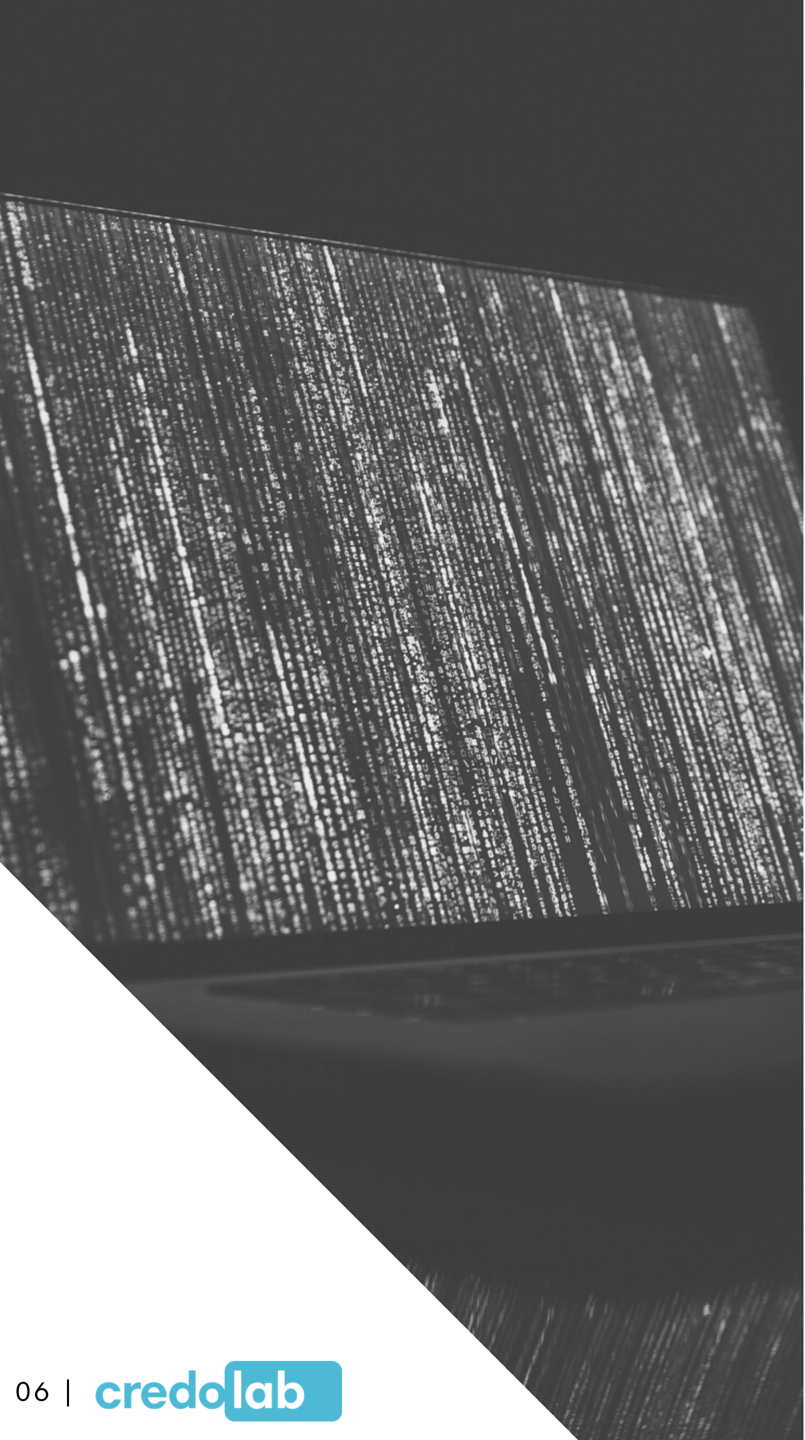
By the use of a simple app, which the users are required to install in their device and give necessary permissions to, the anonymised metadata starts entering the pipeline. The key to factor to take note of here is that the data is 1) not accessed without prior consent by the user, and 2) remains anonymous throughout the process. The data analysis for the purpose of scoring is done only once – at the time of scoring, which takes no more than a few seconds. Once this step is completed, the app does not screen the device for any additional data, nor does it transfer any data out of the phone.

THE SCORING ALGORITHM

Dataset Organization

In order to prioritise and assess model robustness, credolab divides the available data into a training set, a validation set, and a test set. Consistent with recommended practices, the training set is used for fitting the model, the validation set is optionally used to optimise parameters, and the test set is used for reporting the accuracy of the final model with its chosen parameters. Dividing up the dataset in this way for different purposes serves to provide better estimates of actual model performance on individuals who are either new-to-credit or new-to-bank.

credolab's modelling pipeline uses robust machine learning techniques for identifying high-risk borrowers. Given the large number of phone-use features, a number of steps are taken to ensure that (1) the most relevant indicators are picked out for modelling and (2) that the models produce results of comparable quality when applied to unseen data (e.g. a new set of cellphone users).



THE SCORING ALGORITHM

Feature Selection

A range of metrics is used to quantify the information content of each feature, one at a time. Each feature is first discretised via binning and measuring their Information Value. A feature having low Information Value would rank lower in the list of features that can predict the final score.

After excluding features that score low, a second pass through the remaining features looks to eliminate redundant information content. Having reduced the features further, a Random Forest-based approach is used to perform a third pass on further refinement of the set of features. At this stage, the remaining subset of features is ready to be used in predictive modelling.

Information value is a well-established metric used in determining the importance of features in credit scoring (Finlay 2012). Lower Information Values suggest the variable X contains low information about the target, and hence lower in predicting the accurate score.

Random Forests (Breiman 2001) ensembles a random subset of features at each stage of model fitting. Random Forests are traditionally thought to have less risk of overfitting by averaging across a number of decorrelated trees. The use of random features permits the ability to rank the usefulness of features.

THE OUTCOME

The backbone of credolab's modelling engine is the regularised logistic regression. Depending on the stage of the modelling pipeline, it also uses the elastic net logistic regression and the tree-based gradient boosting with grid search - always using out-of-time validation.

The final outcome is a binary indicator of whether the cell phone user becomes in default on their loan. Throughout the data scoring pipeline, we note that demographic information about the individual phone user is not included. Variables such as age, sex, income level, etc. are neither considered for modelling nor extracted from the mobile device for any other purpose.

THE EXPERIMENT

Each part of credolab's modelling pipeline was independently evaluated alongside a range of alternatives using real data. Input features were alternatively considered on both continuous and discrete-binned scales, where applicable. The feature selection and modelling steps were also scrutinised. For example, tree-based approaches such as random forests and gradient boosting via XGBoost (Chen & Guestrin, 2016) were considered as alternatives to logistic regression. 24 other possible versions of the modelling pipeline were considered using some combination of different approaches for feature filtering, feature representation, feature selection, and modelling. The single optimal pipeline that achieved the highest validation set AUC was then compared to that of credolab and achieved test AUC of around 0.70.

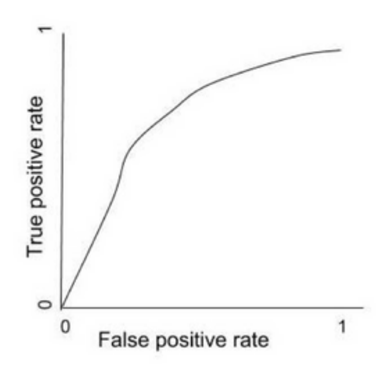
Two different test datasets were considered as part of the assessment: one small dataset (with about 3000 features) and one large dataset (with more than 14,000 features). Test AUC on these datasets were identical between the optimal alternative pipeline and credolab's pipeline, up to the hundredth decimal place.

CONCLUSION

In summary, an independent evaluation of credolab's modelling pipeline was very positive overall. The overarching framework of data splitting, feature selection, model tuning, and model assessment is as statistically sound as it can be. The choices of specific metrics for assessment and feature scoring are appropriate and standard within the domain of application.

When credolab's modelling pipeline was assessed against 24 other possible alternatives that involved using different data splitting, feature selection, and model tuning combinations, the best performing alternative combination did not outperform credolab's approach in test AUC on two different datasets. From this, I conclude that the predictive strength of credolab's modelling approach is comparable to those achieved by other cutting-edge machine learning techniques.

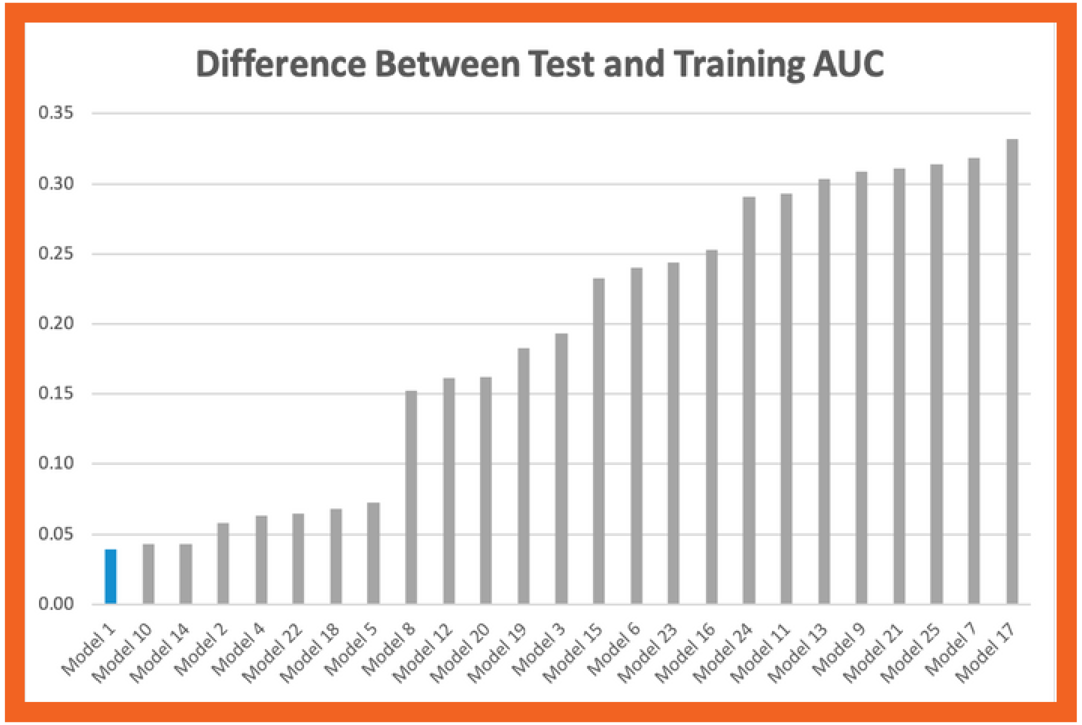
AUC, which stands for Area Under the Receiver Operating Characteristic (ROC) Curve, is used for measuring the predictive performance of any model. AUC has long been an industry standard for scoring classification models in the machine learning community (Hanley and McNeil, 1983). A higher AUC is indicative of better predictive strength.



CONCLUSION

Each model used for this experiment differed in feature filtering, feature representation, feature selection, and modelling. Only credolab's modelling achieved the highest Test AUC value and the least variation in test and training sets - an indicator of a strong scoring model.

The chart below indicates how the different models fared when differences between their test and training AUC were taken into consideration. Credolab's model is marked in blue.



Difference Between Test and Training AUC

ABOUT CREDOLAB

We believe loans improve lives. We also believe traditional banking processes leave a lot of people out of the process. That's why credolab is changing the way the world looks at credit. Our pioneering technology calculates credit scores based on people's mobile and web behavioural data— so lenders can make decisions based on the way people live and work in the modern world.

Making loans more accessible to more people benefits everyone.

Contact us at info@credolab.com to talk more.

Certified fintech.
Recognized by:

SFA
SINGAPORE FINTECH ASSOCIATION

MAS
Monetary Authority
of Singapore

REFERENCES

- Berg, T., Burg, V., Gombović, A., & Puri, M. (2018). On the Rise of FinTechs–Credit Scoring using Digital Footprints (No. w24551). National Bureau of Economic Research.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SigKDD international conference on knowledge discovery and data mining (pp. 785-794). ACM.
- Finlay, S. (2012). Credit scoring, response modelling, and insurance rating: A practical guide to forecasting consumer behaviour. Palgrave Macmillan.
- Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From footprint to evidence: An exploratory study of mining social data for credit scoring. ACM Transactions on the Web (TWEB), 10(4), 22.
- Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148(3), 839-843.
- Sivakumaran, M., & Iacopino, P. (2018). The Mobile Economy 2018. GSMA Intelligence.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

THIS PAPER IS BASED ON THE RESEARCH WORK DONE BY MRS XIAOFEI (SUSAN) WANG, PHD IN OCTOBER 2018

GET IN TOUCH WITH US TO RECEIVE A COPY OF THE FULL STUDY

Email address
info@credolab.com

Website
www.credolab.com