MACHINE LEARNING-DRIVEN CREDIT RISK MODELLING USING SMARTPHONE METADATA

The presentation will explore the highlights of an independent review of CredoLab's scoring model done by Prof. Xiaofei (Susan) Wang, PhD, Lecturer and Associate Research Scholar of the Department of Statistics & Data Science at Yale University. The study explored how CredoLab's proprietary methodology and innovative risk modeling approach identifies the correlation between a user's mobile device metadata and her creditworthiness.

As smartphones facilitate day-to-day interactions such as calling, texting, email, and more, they have become an electronic diary of individual phone users. CredoLab has developed a modeling pipeline that processes this smartphone data in a series of automated steps, rooted in machine learning techniques, and outputs a predictive model for credit default. To protect the confidentiality and to ensure against bias towards individual loan customers, only non-identifying metadata is used.

The independent review Prof. Wang first explored the data sets that CredoLab consumes, how it translates it into scores, and the outcome it serves. She then took a look at how CredoLab's algorithm fared when compared to that of other major players with similar scoring models.

The study found that the backbone of CredoLab's modeling engine is the regularized logistic regression. Depending on the stage of the modeling pipeline, CredoLab 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 smartphone user becomes in default on their loan. Throughout the data scoring pipeline, demographic information such as age, sex, income level, etc. are neither considered for modeling nor extracted from the mobile device for any other purpose. Hence, no bias is built into the model.

Given the very large number of phone-use features, the study found that a number of steps are taken to ensure that (1) the most relevant indicators are picked out for modeling and (2) that the models produce results of comparable quality when applied to unseen data (e.g. a new set of cellphone users).

In summary, when Prof. Wang assessed CredoLab's modeling pipeline 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. The audience will see AI and alternative data sources generating results meant for the financial industry in the digital age, empowering the underbanked and unbanked.