CITY UNIVERSITY OF HONG KONG

香港城市大學

An Innovative Machine Learning Approach to Know-Your-Customer Risk Assessment in Digital and Virtual Banks

以機器學習技術為數字及虛擬銀行評估客戶相關風險的創新方案

Submitted to
College of Business

商學院

in Partial Fulfilment of the Requirements
for the Degree of Doctor of Business Administration

工商管理學博士學位

by

LEUNG Kin Man

梁建文

August 2023

二零二三年八月

Abstract

Know-Your-Customer (KYC) is the first function in an Anti-Money Laundering (AML) regime that is pivotal to banks and financial institutions (FIs) offering services to customers. Although majority of customers are decent and their funds is deemed to be clean, banks and FIs must nonetheless put in place stringent KYC/AML policies and robust checks-and-balances procedures to ensure that customers who come onboard are fit-and-proper, and the bank accounts opened for them will not be used for laundering dirty money from illicit sources.

KYC procedures typically entail mandatory in-person face-to-face (F2F) interviews at physical branches, where identity verification, supporting document and reference checks are performed manually, followed by a due diligence and risk assessment process based on a set of rules and templates. Such KYC procedures often take days or weeks, sometimes months, before any decision can be made.

As financial institutions, particularly the fast-emerging Digital Banks and Virtual Banks (DB/VBs), are adopting electronic platforms to conduct remote non-F2F customer onboarding and instant account opening, the conventional methods of conducting KYC checks are increasingly found to be unsuitable and untenable. Without any in-person onsite interviews, however, the **accuracy** and **adequacy** of the onboarding decision may be poor. Because the rules are subject to different human interpretations, the **consistency** and **explainability** of such a decision cannot be ascertained. And if the process is not completed in minutes, instant account opening becomes impossible.

Moreover, the advent of Open Banking (OB), which enables the sharing of customer and financial data through standard interfaces with non-bank service providers such as Financial Technology (FinTech) start-ups and lifestyle product portals, has aggravated the KYC/AML problem and complicated the underlying process. While these market entrants could introduce new business models, they might also become potential "weakest links" or "loopholes" if they ever adopted unsafe KYC practices and unproven cybersecurity technologies, and as a result, exposed hidden vulnerabilities in the banking infrastructure that could be exploited by fraudsters.

This thesis proposes a Machine Learning (ML) approach to tackling the KYC/AML challenges encountered by financial institutions and particularly DB/VBs. Through training and testing experiments, eleven ML models are evaluated for undertaking the **KYC Risk Assessment and Decision** task, which is the final and indeed most critical decision point in the Customer Due Diligence process, and for predicting the "Accept", "Reject" or "Defer" outcome of each customer onboarding and account opening application. Most significantly, riding on the evaluation results,

an innovative Balanced Scorecard Template and a Dual/Multi-ML Models Framework are devised to meet the **accuracy** and **explainability** needs of the industry regulators and stakeholders.

To facilitate training and testing of the ML models, a KYC-Sim Dataset has been composed with over 1000 natural and synthetic datapoints, with decisions tagged by KYC experts of two data source banks. It includes five categories of **feature variables** (i.e., Identity and Demographic, Credit and Financial, Business and Professional, Behavioural and Transactional, Social and Reputational) that collectively determine the KYC risk assessment outcomes, along with **moderating variables** (i.e., Data, Technology, Process, Expertise, Laws and Regulations) that influence the decision process. The Dataset can be enhanced and enriched in production use over time, and will become more instrumental for research projects in KYC/AML areas going forward.

The experimental results are analysed quantitatively and discussed qualitatively to demonstrate that many of the proposed ML models, such as Artificial Neural Network, Random Forest, and Stochastic Gradient Descent classifiers, can achieve an over 90 percent **accuracy** on predictions of “Accept” and “Reject” outcome classes**,** as measured by weighted average F1-Scores, and around 80 percent in the case of “Defer” class for further investigation. This automated ML approach also ensures higher **consistency** and **efficiency** of prediction than conventional rule-based methods that are conducted largely manually.

This research assesses and assigns ratings on the **explainability** aspect of the ML models evaluated. It demonstrates that the Balanced Scorecard Template, as mentioned above, can facilitate users to adopt the most suitable ML model, whereas the Dual-ML Models Framework to select an appropriate combination of ML classifiers, depending on the relative emphases or weights assigned to accuracy and explainability respectively in response to the circumstances and needs they face from time to time.

The proposed ML approach to **KYC Risk Assessment and Decision** is a pioneering research work that provides DB/VBs and FIs, as well as their corporate auditors and industry regulators, clear argument, solid evidence, and strong confidence on the use of ML approach and methodologies for KYC risk assessment process. It also bears great relevance and wide application for many other digitally enabled businesses and industries that render instant customer onboarding and servicing in a remote non-F2F manner as well.