ISO/IEC JTC 1/SC 42/SG 3

Use cases and applications

Convenorship: JISC (Japan)

Document type: National Body Contribution

Title: 027_ISO-IECJTC1-SC42-SG3 NXXXX India NB Contribution Credit ScoringvF

Status:

Date of document: 2018-08-17

Expected action: INFO

Email of convenor: maruyama.f@jp.fujitsu.com

Committee URL: https://isotc.iso.org/livelink/livelink/open/jtc1sc42sg3

General

ID[1]				
Use case name	Credit scoring using KYC data			
Context	Banking and Financial Services			
Application domain	On-premise systems			
Status	PoC			
Contributor	Name	Affiliation	Contact	
	Rohit Pandharkar	Mahindra Group	PANDHARKAR.ROHIT@mahindra.com	
Scope[2]	Building a risk scorecard for loan applicants using KYC data for better risk management and high population coverage			
Objective(s)	Assigning a risk score to every loan applicant in real time, using just KYC data, which will ensure both new-to-credit and mature customers can be assessed for their creditworthiness, and offered loans on appropriate terms			
Narrative	Short description (not more than 150 words)	It can be often difficult to build a risk scorecard using only KYC data, which often has noisiness and incompleteness issues. However if realized, it can be used to provide a objective score to all loan applicants, even the new-to-credit ones. Non-linear classification algorithms are suitable for this purpose. Several variables are collected from the customer during the KYC process such as Age of customer, Self-reported income, Type of Occupation, Purpose of loan, etc. All these features can be added to a non-linear risk model and their complex interactions allowed to take place.		
	Complete description	Financial institutions find it much easier to assess customers with an existing credit history, or those living in urban areas. There are also several credit bureaus who assist them in this endeavor. However, these frameworks don't work as well for new-to-credit customers, especially in rural areas. If only industry wide models or simple heuristics are used to score such customers, many deserving loan applicants will end up not getting a loan or not getting it at deserving terms. Instead, if a good risk scorecard is built using KYC data, which is collected from every loan applicant as a routine and regulated process, it will ensure every applicant receives an objective score. To tackle this problem, non-linear models such as Random Forest and XGBoost are being used which can accommodate many parameters, including categorical ones, and are reasonably resistant to noise in the data.		
Key performance indicators (KPIs)	ID	Name	Description	Reference to mentioned use case objectives

	1	Delinquency Rate	Percentage of loan defaulters in first X months from loan disbursal vs score bins	Large monotonous decrease in delinquency rate as creditworthiness score increases is desirable, and indicates a good scorecard
	2	Approval rate	Ratio of loan disbursals to loan applicants	Larger approval rate at a predetermined risk level is desirable and indicates a good scorecard

AI features	Tasks(s)	Credit Scoring
	Method(s)[3]	Random Forest, XGBoost and Ensemble models
	Hardware[4]	64 GB RAM, Intel Core i5
	Terms and concepts used[5]	Classification, Bagging, Boosting, Ensembles
Challenges and issues	1. KYC data obtained from extreme rural areas can be noisy, may have several missing values, and needs appropriate preprocessing and treatment before feeding to the model algorithm 2. Non-linear models like Random Forest and XGBoost need significant computational power during the training phase	
Societal concerns	We don't see any societal concerns if it is used	

Data (optional)

Data characteristics	
Description	Historical KYC data available in internal systems
Source[6]	EDW (Enterprise Data Warehouses)
Type[7]	Structured Data
Volume (size)	10 GB
Velocity (e.g. real time)[8]	One-time data dump during training phase, real time in production phase
Variety (multiple datasets)[9]	Mostly Structured
Variability (rate of change)[10]	Moderate
Quality[11]	Moderate

References

References						
No.	Type	Reference	Status	Impact on use case	Originator/organization	Link
1	Paper	[Breiman 01] Leo Breiman. “Random Forests”. Machine Learning, Volume 45, Issue 1, Pages 5-32. 2001.	Published	High	University of California, Berkeley	https://dl.acm.org/citation.cfm?id=570182
2	Paper	[Chen 16]. Tianqi Chen. “XGBoost: A Scalable Tree Boosting System”. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794. 2016.	Published	High	University OF Washington, Seattle	https://dl.acm.org/citation.cfm?id=2939785
3	Paper	[Opitz 99]. David Opitz. “Popular ensemble methods: an empirical study”. Journal of Artificial Intelligence Research. Volume 11 Issue 1, Pages 169-198. 1999.	Published	High	University Of Montana, Missoula, MT	https://dl.acm.org/citation.cfm?id=3013549

Footnote

[1] Leave this cell blank.

[2] The scope defines the limits of the use case.

[3] AI method(s)/framework(s) used.

[4] Hardware system used.

[5] Terms and concepts listed here can be used to extend the work of WG 1 (AWI 22989 and AWI 23053) as necessary.

[6] Origin of data, which could be from instruments, IoT, web, surveys, commercial activity, or from simulations.

[7] Structured/unstructured Images, voices, text, gene sequences, and numerical. Composite: time-series, graph-structured

[8] The rate of flow at which the data is created, stored, analysed, or visualized.

[9] Data from a number of domains and a number of data types. The wider range of data formats, logical models, timescales, and semantics complicates the integration of the variety of data.

[10] Changes in data rate, format/structure, semantics, and/or quality.

[11] Completeness and accuracy of the data with respect to semantic content as well as syntactical of the data (such as presence of missing fields or incorrect values)