RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Audit with Machine Learning: Applying an Unsupervised Algorithm on General Ledgers of an Australian Bank

Kathy (Danyang) Wei, Lanxin Jiang, and Yu Gu

Miklos Varsarhelyi, Soo Hyun Cho

11/8/2020

Background

- Researchers and practitioners have proved the ability of machine learning to learn data patterns and have applied it to different contexts in the accounting field:
  - Predicting accounting fraud (Perols, 2011; Perols et al., 2017)
  - Investigating the prediction of corporate bankruptcies or defaults (Barboza et al., 2017)
  - Improving accounting estimates (Ding et al., 2020)
- This research aims to exhibit the potential of machine learning to identify and determine true outlying transactions

Concerns from Auditors and Solutions

Concerns

- "Black box" issue
- Indirect output
- Work overload (i.e., false positives)

Solutions

- Data reconstruction
- A distance-based, unsupervised algorithm
- A framework

Data (FY 2019)

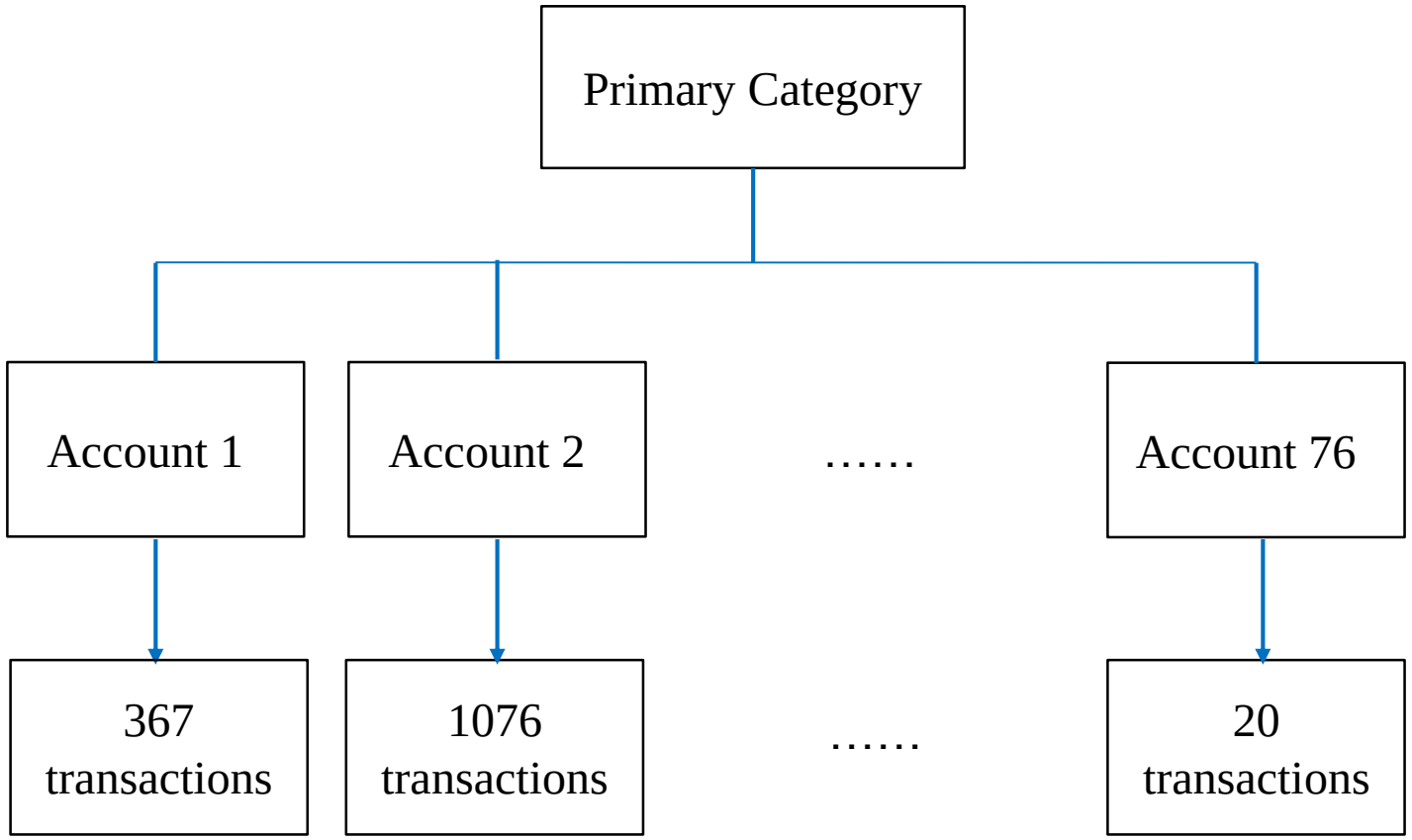
- Trial balance (total: 1,647 accounts) → Account level
- General ledgers (total: 1,438,790 records) → Transaction level

Stateme	Account Numbe	Description	Primary	Current Period	Prior Period
BS	10.1014.10107	NAB MAIN ACCOUNT (50-922-3645)	Cash and Cash Equivalents	3,081,471.20	10,541,489.73
BS	10.1014.10115	ATM CASH FLOATS	Cash and Cash Equivalents	640,780.00	838,480.00
BS	10.1014.10201	CENTRAL BORROWING AUTHORITIES PROMISSORY NOTES	Cash and Cash Equivalents	34,883,097.45	-
BS	10.1014.10220	11AM CALL ACCOUNT - BANKS	Cash and Cash Equivalents	83,500,000.01	7,000,000.01
BS	10.1014.10235	FLOATING TERM DEPOSITS - OTHER ADIS	Cash and Cash Equivalents	52,032,505.00	52,032,505.00
BS	10.1014.10238	CENTRAL BORROWING AUTHORITIES 11AM CALL ACCOUNT	Cash and Cash Equivalents	82,000,000.00	-
BS	10.1014.31024	OVERNIGHT ACCOUNT ACCURED INTEREST	Cash and Cash Equivalents	(5,136.99)	-
BS	10.1014.31028	RESIDENT FINANCIAL INSTITUTIONS - ACCRUALS	Cash and Cash Equivalents	(4,535.33)	-
BS	10.1014.31301	11AM CALL ACCOUNTS ISSUED BAL	Cash and Cash Equivalents	(50,000,000.00)	-
BS	10.2400.10105	PETTY			
BS	10.3100.10105	PETTY			
BS	10.3200.10105	PETTY			
BS	10.3410.10104	CASHI			
BS	10.3410.10105	PETTY			
BS	10.3420.10104	CASHI			

Account Code	Transaction Id	Net	Effective Date	Created Date	Document Type	User Id	Reference	Journal Description	Line Description
10.1011.30944	PCARD	88	6/6/18	17/09/2018	PCSTAT	ALDREDJ		AICD Cyber Resilience	
10.1011.30944	PCARD	565.59	11/6/18	17/09/2018	PCSTAT	ALDREDJ		Annual Renewal to AICD	
10.1011.30944	PCARD	1951.44	26/06/2018	17/09/2018	PCSTAT	ALDREDJ		Registration to WOCCU Singapore	
10.1011.30944	PCARD	2150.74	26/06/2018	17/09/2018	PCSTAT	ALDREDJ		Accommodation re WOCCU Singapore	
10.1011.30944	PCARD	22.38	12/7/18	17/09/2018	PCSTAT	ZANCOP		WOCCU - Taxi Singapore Airport	
10.1011.30944	PCARD	19.28	12/7/18	17/09/2018	PCSTAT	ZANCOP		WOCCU - Meal	
10.1011.30944	PCARD	40.01	13/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU Meal	
10.1011.30944	PCARD	6.99	13/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU - Meal	
10.1011.30944	PCARD	4.03	14/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU Meal	
10.1011.30944	PCARD	38.36	17/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU Meal	
10.1011.30944	PCARD	54.61	18/07/2018	17/09/2018	PCSTAT	ZANCOP		meal WOCCU trip	
10.1011.30944	PCARD	17.99	19/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU - Meal	
10.1011.30944	PCARD	226.79	19/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU - Meal	
10.1011.30944	PCARD	73.78	19/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU Meal	
10.1011.30944	PCARD	143.97	20/07/2018	17/09/2018	PCSTAT	ZANCOP		WOCCU - Taxi Airport to Home	
10.1011.30948	PCARD	2877.94	19/07/2018	29/08/2018	PCSTAT	ZANCOP		WOCCU Accommodation	

General Ledgers

Data

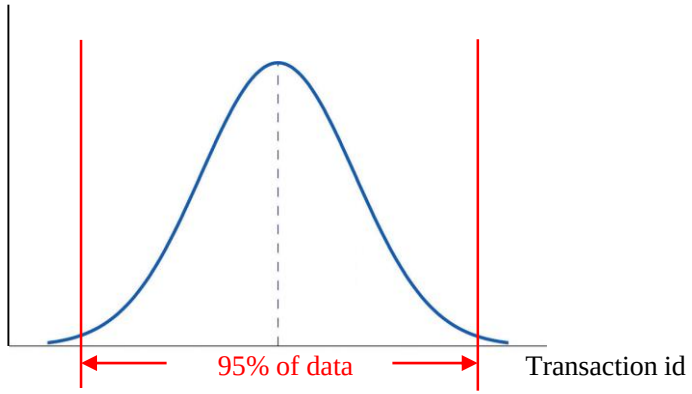

Primary Category
Account 1
Account 2
……
Account 76
367 transactions
1076 transactions
……
20 transactions

Data Reconstruction

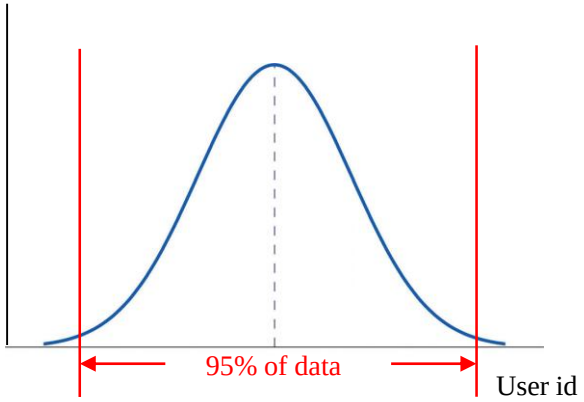
- Account level (9 variables)

Variable Name
Mean (positive)
Standard deviation (positive)
Mean (negative)
Standard deviation (negative)
Account balance change
Percentage of abnormal transactions based on transaction id
Percentage of abnormal transactions based on user id
Mean (days)
Standard deviation (days)

Count of each unique transaction id



Count of transactions entered by each unique user id



Data Reconstruction

- Transaction level (5 variables)

Variable Name
Standard score (net amount)
Net amount frequency
Transaction id frequency
User id frequency
Standard score (days)

Algorithm: Minimum Covariance Determinant

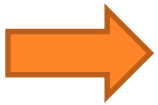
- A distance-based unsupervised algorithm
- Consider all observations in the population as one cluster
- A center is first found and then the distance of each point to the center will be calculated
- Any observation that has a distance above a certain cutoff value is treated as an outlier
- Mahalanobis distance is used for distance calculation and the correlation between variables are included

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

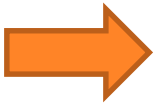
Example – Transaction Level

	standard score (net)	frequency (net)	frequency (transaction id)	frequency (user id)	standard score (days)
0	-0.064130	0.002129	0.001588	0.002502	-1.370394
1	-0.064108	0.001456	0.001588	0.002502	-1.370394
2	-0.063994	0.000004	0.001588	0.002502	-1.370394
3	-0.063974	0.000633	0.001588	0.002502	-1.370394
4	-0.063970	0.000042	0.001588	0.002502	-1.370394

Transaction Level Input (Example Category: Loans and Advances to Members)



Minimum Covariance Determinant



	Account Code	Transaction Id	Net	Effective date	Created date / time	Document Type	User Id	Journal Description	Mahalanobis
0	10.1014.10410	44861	-50.00	2019-05-21	2019-05-22	APJNL	CHANDLERC	NaN	84541.014010
1	10.1014.10418	44861	-100.00	2019-05-21	2019-05-22	APJNL	CHANDLERC	NaN	39479.180237
3	10.1014.10438	44861	400.00	2019-05-21	2019-05-22	APJNL	CHANDLERC	NaN	7417.799210
4	10.1014.10448	44861	-408.97	2019-05-21	2019-05-22	APJNL	CHANDLERC	NaN	35.286840
5	10.1014.10410	44861	500.00	2019-05-21	2019-05-22	APJNL	CHANDLERC	NaN	8795.467036

Transaction Level Output (Example Category: Loans and Advances to Members)

Example – Account Level

	Mean(Positive)	STD(Positive)	Mean(Negative)	STD(Negative)	AccountBalanceChange	Abnormal Transaction ID percent	Abnormal User ID percent	Means(days)	STD(days)
0	3598565.35	15347111.54	-4185824.56	15798463.19	7460018.53	0.004620	0.015400	181.95	105.41
1	117191.30	163248.86	-98043.86	147376.65	197700.00	0.038835	0.097087	179.71	106.90
2	14244154.87	15886333.19	-27777777.78	12774758.10	34883097.45	0.137931	0.034483	133.38	104.94
3	26092233.01	24684057.29	-19631578.95	17244992.11	76500000.00	0.101695	0.025424	193.08	104.00
4	27900000.00	15015430.66	-19040000.00	7971198.15	82000000.00	0.066667	0.044444	54.38	43.23

Account Level Input (Example Category: Loans and Advances to Members)

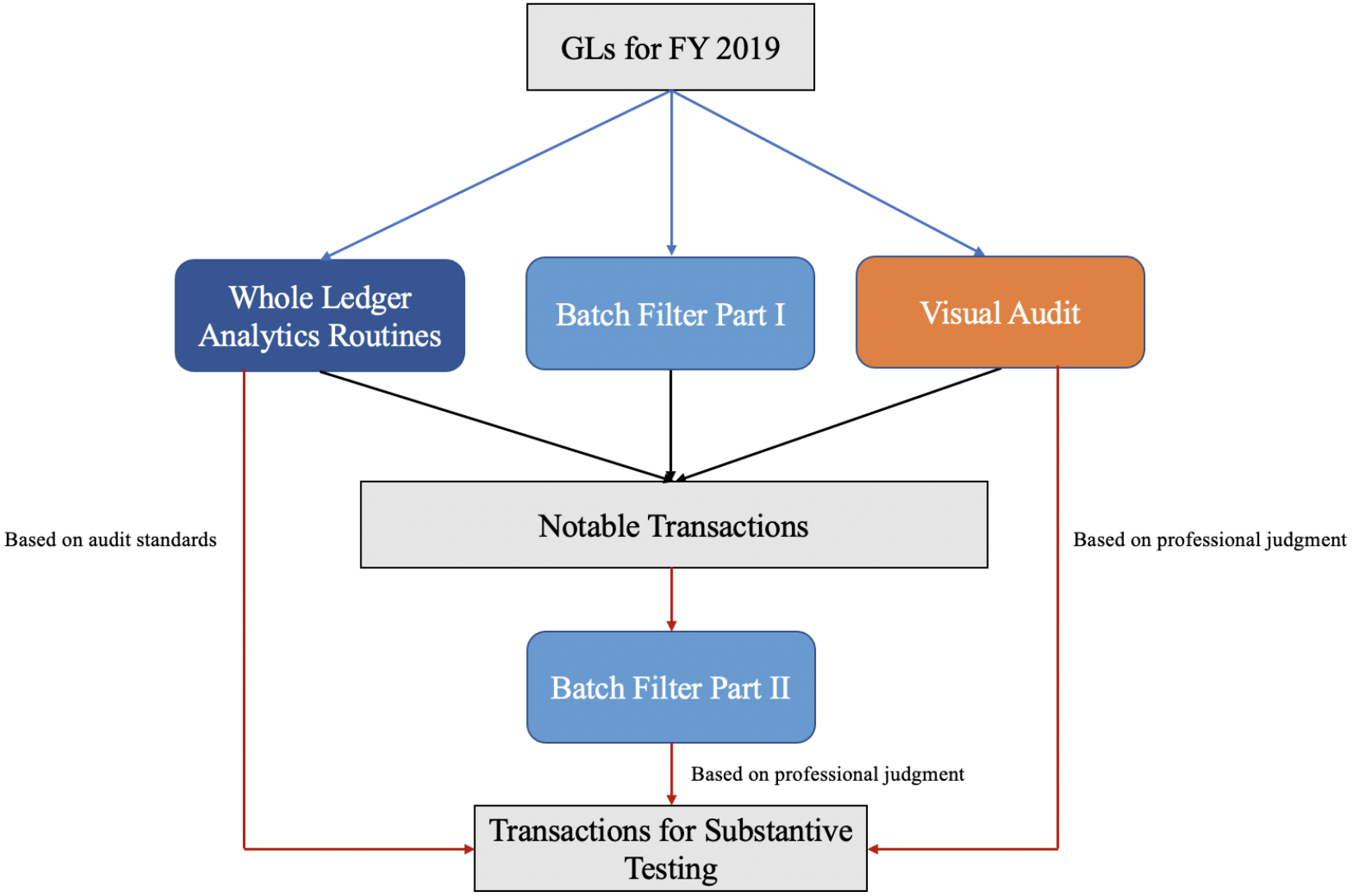
Minimum Covariance Determinant

	Account	Mahalanobis	Count	Total Transactions	Percentage
0	10.1014.10404	5.409930	25	58	0.43
1	10.1014.10405	830.793204	124	755	0.16
2	10.1014.10407	3.059699	212	728	0.29
3	10.1014.10410	3.761143	246	415	0.59
4	10.1014.10412	2409.234612	215	790	0.27

Account Level Output (Example Category: Loans and Advances to Members)

Batch Filter

- Each primary category is considered as a batch
- The batch filter consists of two parts that have different functions
  - Part I (account level): Identifying notable transactions among whole population
  - Part II (transaction level): Identifying transactions for substantive testing among the notable transactions


GLs for FY 2019
Whole Ledger Analytics Routines
Batch Filter Part I
Visual Audit
Notable Transactions
Based on audit standards
Based on professional judgment
Batch Filter Part II
Based on professional judgment
Transactions for Substantive Testing

Result

- Validation dataset: A list of transactions that went to substantive testing
- The batch filter identifies 90% of transactions in the list
- For some categories, false positives are high

THANK YOU!