102nd European Study Group with Industry: Corlytics.

Team Members: James Sweeney, Damien McParland, Arthur White, Adrian O’ Hagan, David O’Sullivan, Davide Cellai, Jan Idziak, Alessandro Montagnani

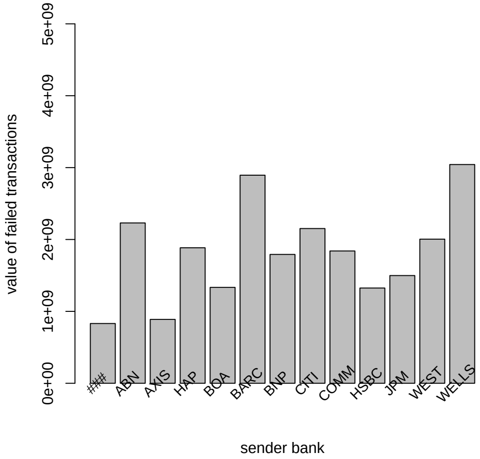
Overview.

- Data
- Summary Statistics.
- Clustering Results.
- Further Work: association rules, classification trees, logistic regression.

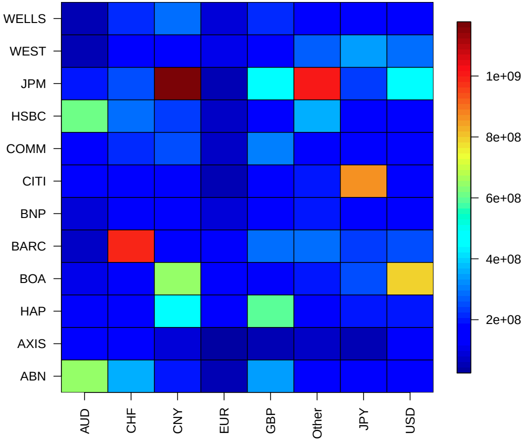
Data.

- Data provided for 59527 **failed** financial transactions on 15 consecutive days.
- Complete data available for 58640 of the transactions.
- Main variables of interest:
  - share price
  - share quantity
  - issuing bank
  - receiving bank
  - currency
  - failure value (price × quantity)
- Examined data for trends in transaction failures and combinations of high failure value intensity.

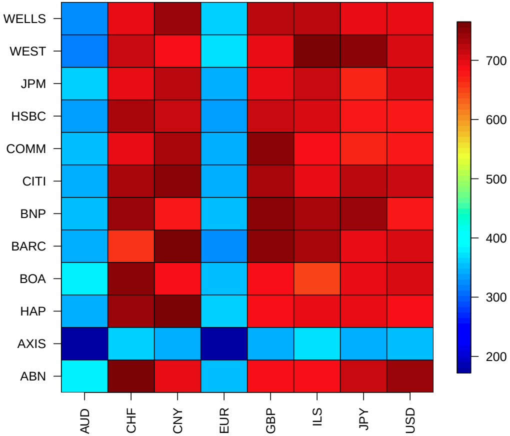
Bar Chart of Total Value of Failed Transactions by Bank



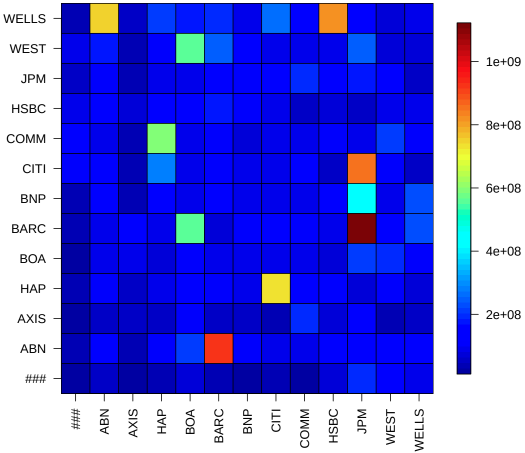
Heat Map of Bank to Currency by Failure Value Intensity



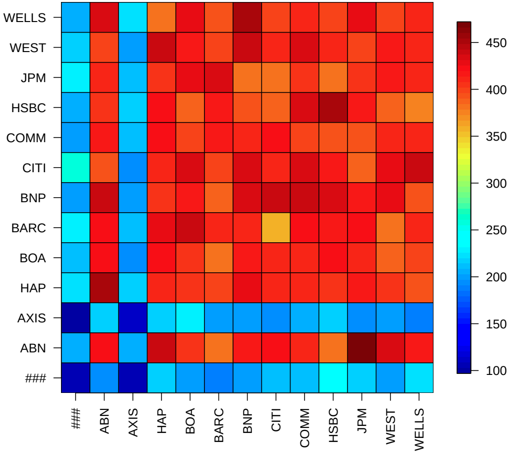
Heat Map of Bank to Currency by Number of Failures Intensity



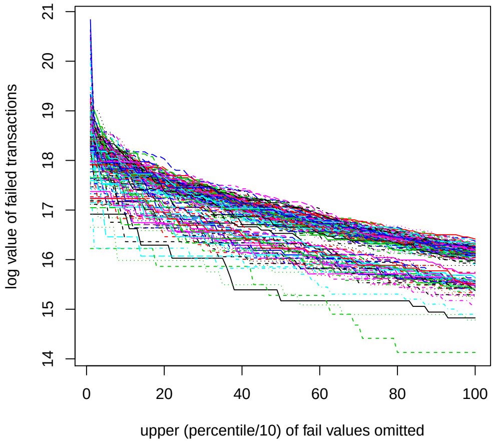
Heat Map of Bank to Bank by Failure Value Intensity



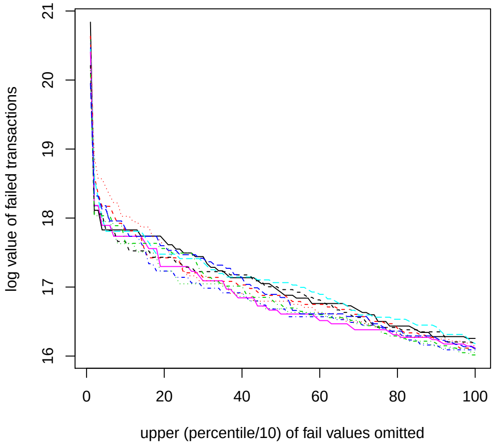
Heat Map of Bank to Bank by Number of Failures Intensity



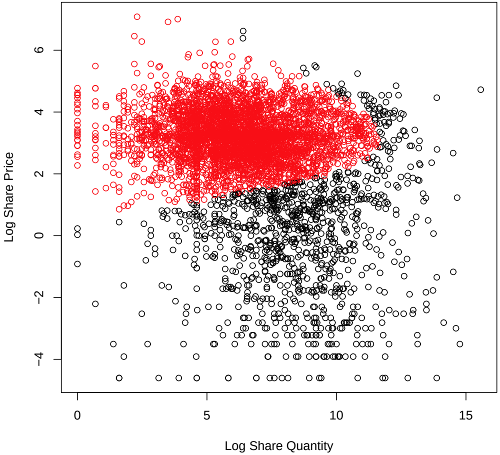
Scree plot of total value of fail transactions versus percentile of fail values omitted by bank pairing



Scree plot of total value of fail transactions versus percentile of fail values omitted: worst bank pairings



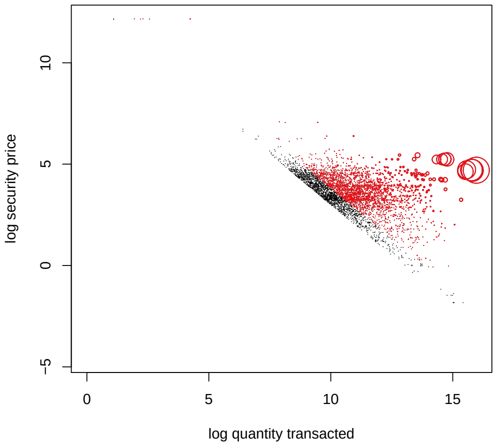
Mixture of 2 skew-normal distributions for log share price versus log share quantity



Extreme fails.

- The data is dominated by the extreme fail outcomes.
- From a total fail value across all 59527 transactions:
    - Over 30% due to the 67 most extreme fail transactions.
    - Over 50% due to the 172 most extreme fails.

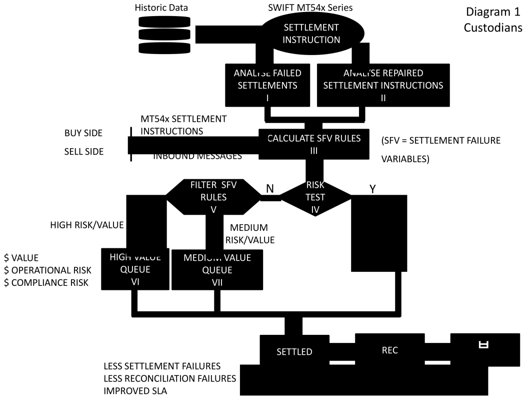
Scatter plot of log share security price versus log quantity scaled by size of fail value



Further Work

- Receive data for **successful** financial transactions on same days.
- Re-run previous analyses and compare results versus the failed transactions.
- Road map of potential strategies: 3 methods coded and applied to German Credit Data for demonstrative purposes.

The General Problem: Solutions Flowchart



Historic Data

SWIFT MT54x Series

SETTLEMENT INSTRUCTION

Diagram 1
Custodians

ANALYSE FAILED SETTLEMENTS
I

ANALYSE REPAIRED SETTLEMENT INSTRUCTIONS
II

MT54x SETTLEMENT INSTRUCTIONS

BUY SIDE

SELL SIDE

INBOUND MESSAGES

CALCULATE SFV RULES
III

(SFV = SETTLEMENT FAILURE VARIABLES)

FILTER SFV RULES
V

N

RISK TEST
IV

Y

HIGH RISK/VALUE

MEDIUM RISK/VALUE

$ VALUE
$ OPERATIONAL RISK
$ COMPLIANCE RISK

HIGH VALUE QUEUE
VI

MEDIUM VALUE QUEUE
VII

SETTLED

REC

LESS SETTLEMENT FAILURES
LESS RECONCILIATION FAILURES
IMPROVED SLA

German Credit Data.

- Data for 1000 loan applications to a German Bank (available on the UCI clustering database).
  70% granted and 30% not granted.
- Each observation contains information for:

  Loan Installment Rate
  Number of Debtors
  Type of Property Owned
  Other Loan Installment Plans
  Number of Existing Loans
  Maintenance Payments Status
  Foreign Worker Status
  Duration as Resident
  Purpose of Loan
  Savings Bonds Indicator

  Applicant Marital Status
  Duration of Loan
  Age
  Housing Status
  Job
  Telephone Code
  Checking Account Indicator
  Credit History Rating
  Amount
  Employment Status

- **Loan Granted Indicator**.

German Credit Data

- There is also an associated cost matrix for predictions for loan granted indicator.
- Prediction of "loan granted" where not granted in data incurs 5 times the cost of prediction of "loan not granted" where granted in data.
- Zero cost for predicting "loan granted" and "loan not granted" where granted and not granted respectively.
- This is analogous to the "cost of failed transaction" and "fine for failed transaction" features of the Corlytics data.

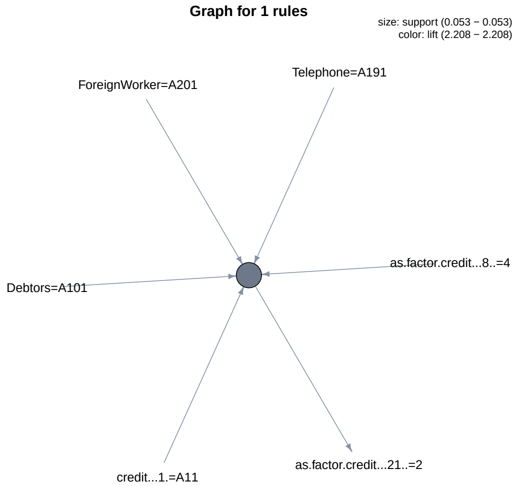
Potential Strategy 1: Association Rules

- Apply association rules for mixed fail/success transaction data to uncover relationships between the variables
- Include outcome of transaction (fail/success) as a variable. Ordering of variables within the transaction unimportant.
- Automated via the **arules** package in **R**.
- Continuous variables converted to binary form and factor variables retained at original number of levels.

Potential Strategy 1: Association Rules for German Credit Data

- Initial model, using all available variables, suggests over 10000 association rules.
- Model recalibrated using lift and confidence to sort rules, isolating optimal combinations.
- Rule with highest lift based on application fulfilling:
  checking account with negative balance
  quarterly installment rate
  no guarantors
  no telephone number
  foreign worker
- Lift 2.2: application 2.2 more times likely to fail if it obeys the specified rule.
- Probability of loan failing under the specified rule: 0.66. Applies to 53 customers in dataset.

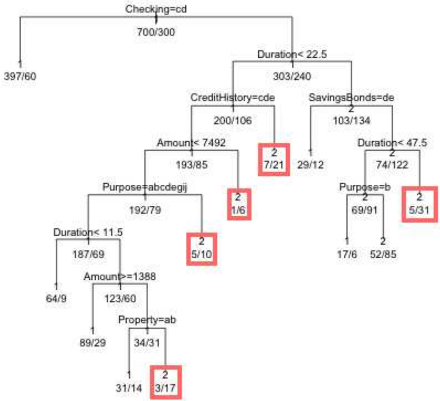
Potential Strategy 1: Association Rules for German Credit Data



Potential Strategy 2: Classification Trees

- Data mining method. Predict value of target variable (transaction pass/fail) based on input variables cast as a series of nodes with finite stem outcomes.
- Greedy algorithm for top-down learning for the model.
- White-box: easy to understand in the context of the data versus commonly employed alternatives such as artificial neural networks.
- Robust, relatively fast to train for large data and can be validated using standard statistical tests.

Potential Strategy 2: Classification Trees for German Credit Data



Potential Strategy 3: Logistic Regression

- Letting *p* be the probability of a transaction failing:

$$\log(p/1-p) = \alpha + \beta_1 * log_{price} + \beta_2 * log_{quantity}$$

$$+\gamma_{iBank} + \gamma_{rBank} + \gamma_{iBank*rBank} + \cdots$$

- Maximum likelihood estimates of parameters from *glm* package in **R** or similar.
- Sensitivity analysis of threshold value p^* such that a transaction should be flagged as a potential fail (not necessarily 0.5, potentially 0.05).
- Can use individual *p* values to infer expected fine cost for all transactions.

Potential Strategy 3: Logistic Regression applied to German Credit Data

- Training data: 600 randomly selected observations.
  Test data: the 400 remaining observations.
- Classification accuracy on test data for a model using all variables (without polynomial or interaction terms): 75%.
- Misclassification w.r.t. cost: 0.98, improvement from 1.242 from null model.