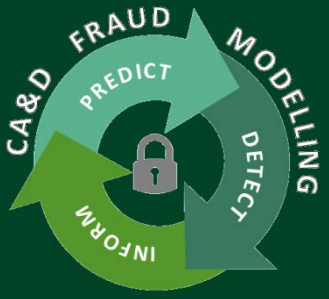
A SURVEY OF RANDOM FOREST USAGE FOR FRAUD DETECTION AT LLOYDS BANKING GROUP

Adam Langron

AUGUST 2015



LLOYDS BANK

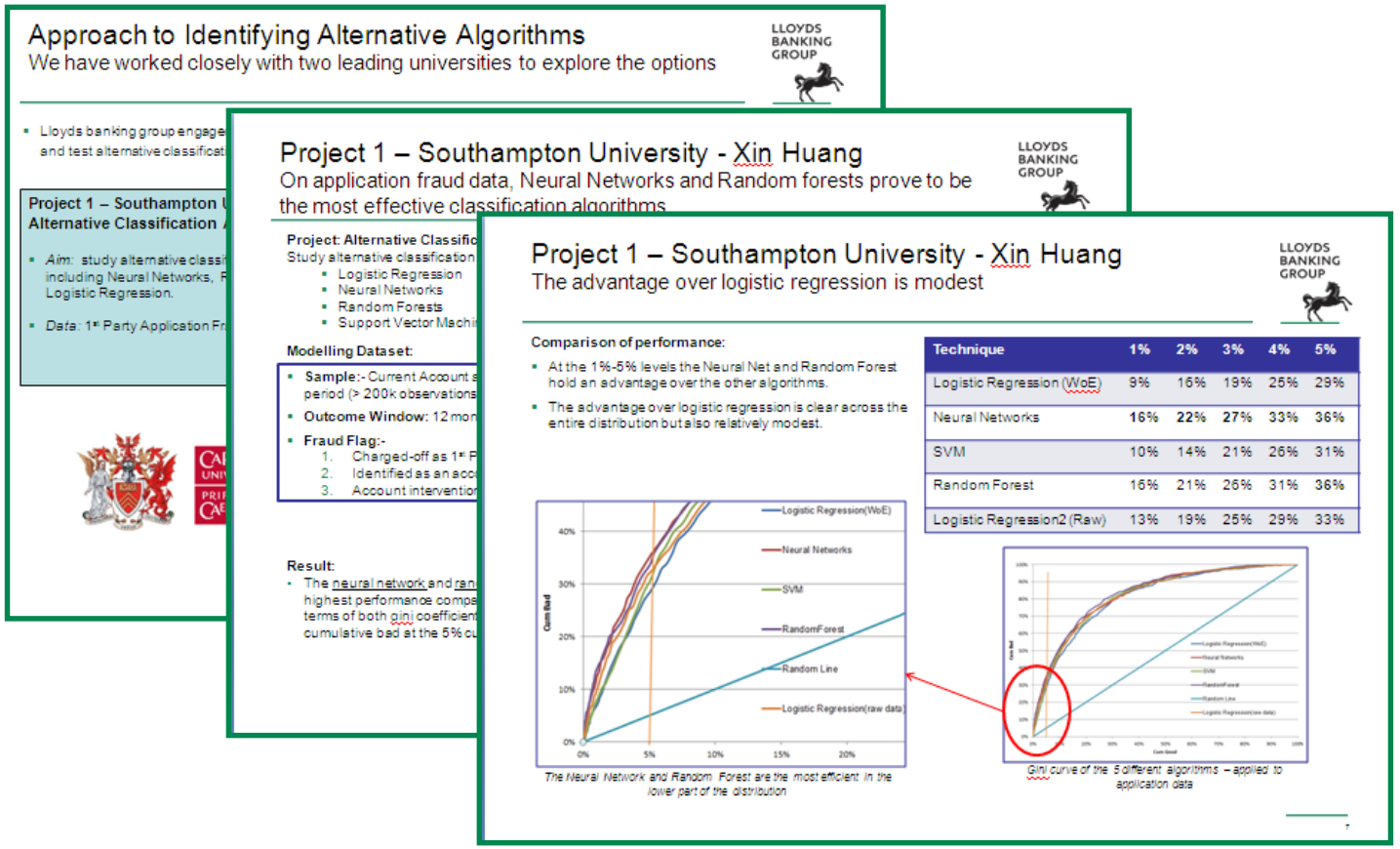
CONTENTS

BACKGROUND

BACKGROUND

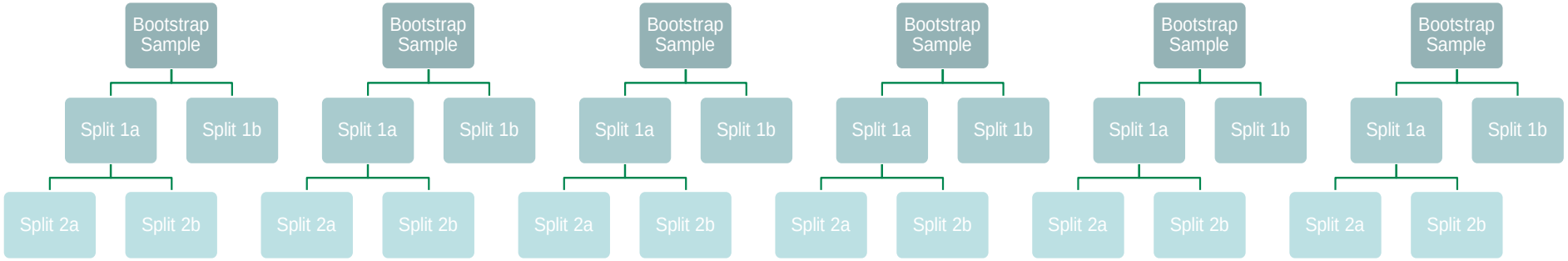
2012 Masters project with Xin Huang suggested that random forests would provide improved discrimination over traditional logistic regression models.

This formed the basis of 2013 CRC conference paper by Kevin Barrett.



RANDOM FORESTS

BRIEF INTRODUCTION



Hundreds of trees – a forest!

These leaves at the bottom of the trees ‘vote’

- If > 50% of the training sample were good at this leaf, the vote is ‘good’
- If > 50% of the training sample were bad at this leaf, the vote is ‘bad’

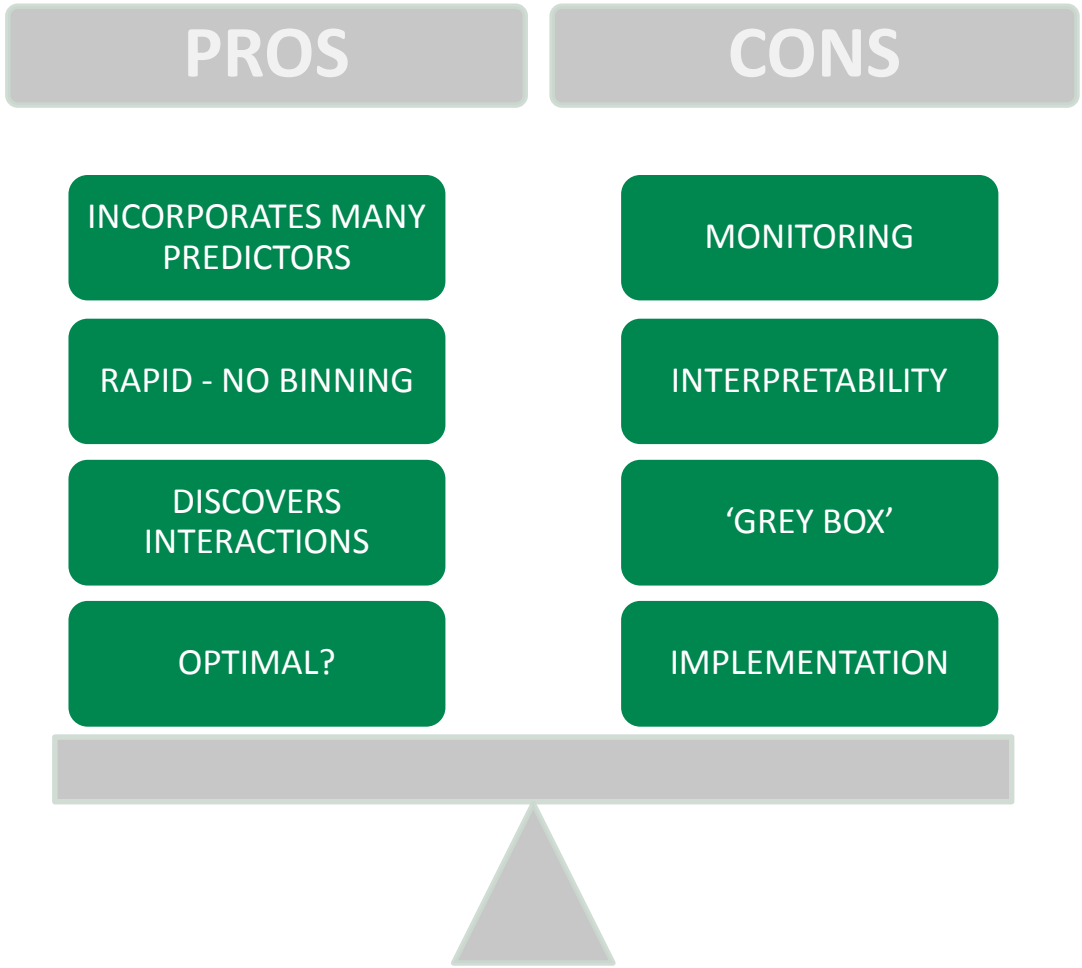
Each new observation will fall to one final leaf node in each tree

Votes are counted and a predicted outcome assigned

- This is a ratio that can be interpreted as a score

RANDOM FORESTS

PROS AND CONS

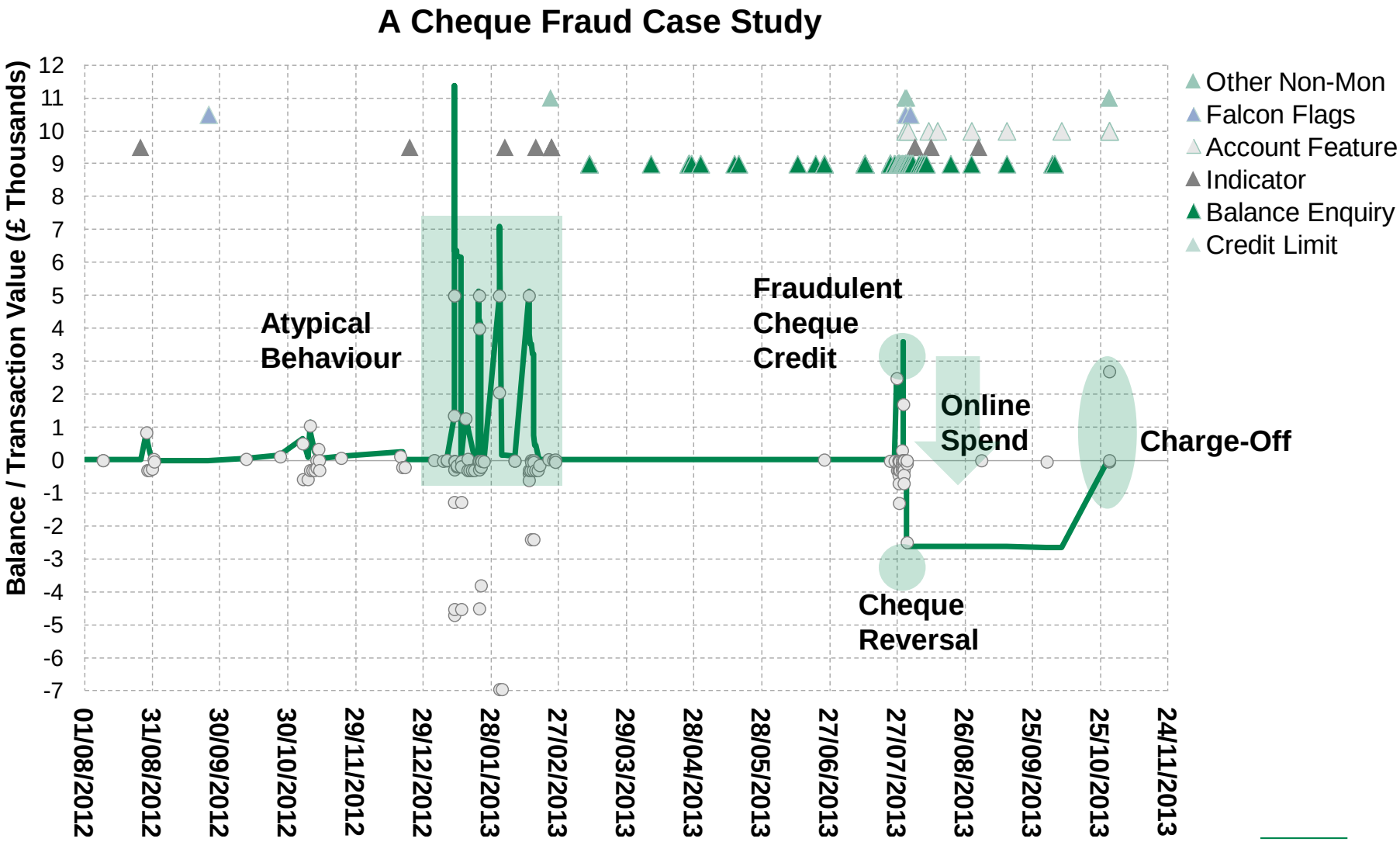


CURRENT ACCOUNT FRAUD

FIRST PARTY FRAUD
DETECTION USING PERSONAL
CURRENT ACCOUNT
TRANSATIONS

CURRENT ACCOUNTS – THE PROBLEM

What is the best way to use the vast amount of data that the bank holds to detect fraud and refer accounts **at the right time** to allow prevention?



CHARACTERISTICS

The richness of transactional data allows over 100,000 possible predictors to be created

Point of Sale
- Visa Debit authorisations
- Accepts/Declines
- Grouped by Merchant Category Code

Counter
- Branch transactions
- Deposits and withdrawals

Transfers
- Non–counter payments and transfers
- Faster Payments, DD
- All channels

Enquiries
- Non-monetary events
- Balance enquiries
- Account status checks
- Grouped by channel

X

Days
- Days from obs point
- 0 -10 days

Weeks
- Weeks from obs point
- 0 - 6 weeks

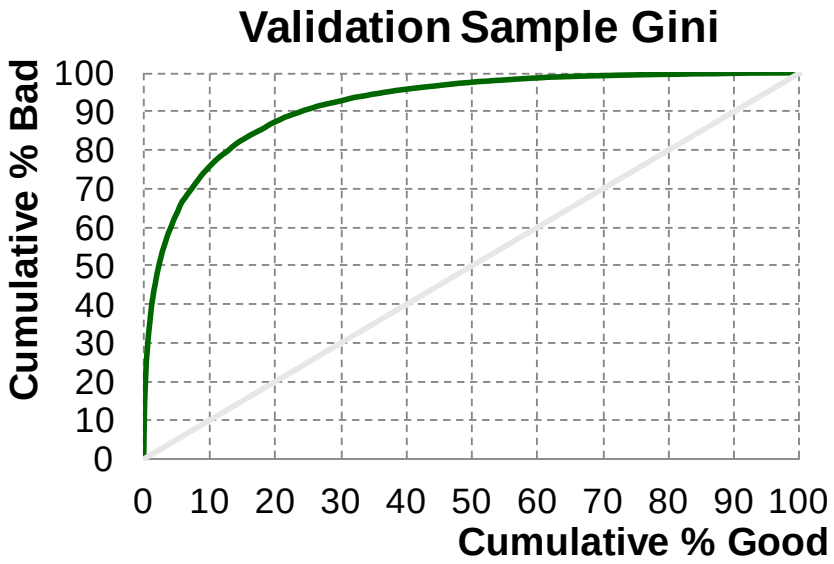
X

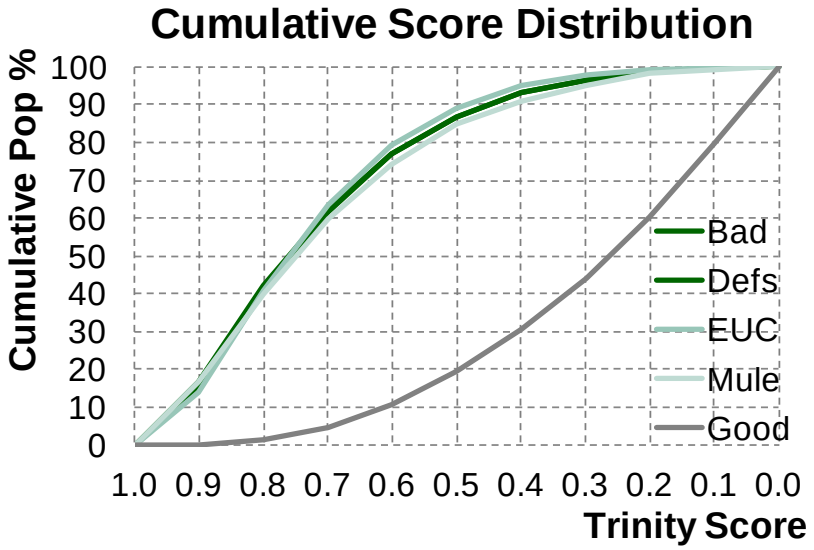
Measure
- Volume
- Value
- Accept/Decline
- Excess
- Utilisation

- Combining transaction type with time interval and measure yields a multitude of predictors
- Using Merchant Category Code and Money Manger within the relevant transaction types enabled even more granularity
- Infrastructure necessitated the reduction of these characteristics
  - Information Value macros cannot cope with this number of chars
  - Chars populated very sparsely were removed
  - Chars with suspiciously high bad rates were removed if there was evidence of fraud investigation activity
  - Information Value and was used to select final candidate chars

MODEL PERFORMANCE

Model performance on the validation sample is strong



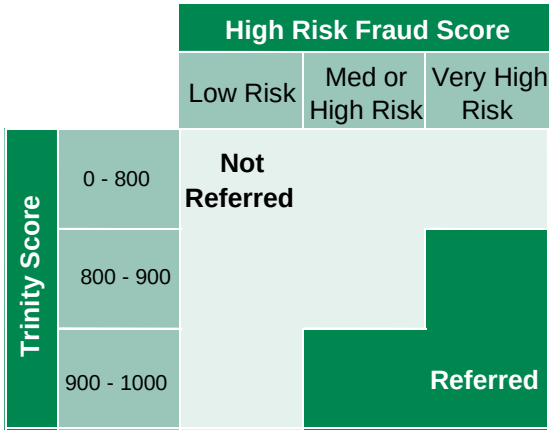


- Performance is shown for the development validation data set
- Discrimination is very high: Gini is in excess of **80%**
- Validation has been performed on three out of sample holdouts and one out of time holdout and shows stable performance
- The model discriminates well for all three fraud categories: First Party Definitions, EUCs and Mules
- Investigator feedback from a small sample verified that the model ranks effectively and finds a wide variety of fraud typologies

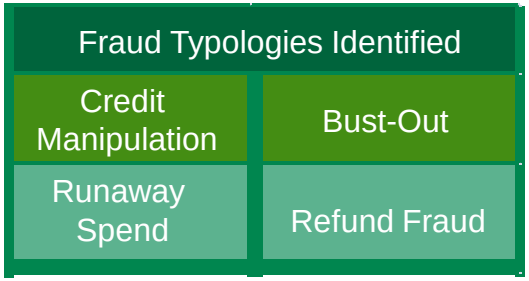
IMPLEMENTATION

Accounts now being referred to Fraud Operations

- *The model was initially piloted in an Operational environment*
  - To confirm referrals were generated at the 'right' point
  - That volumes were manageable
  - That Trinity could 'add value' to the existing strategy suite
- *Pilot results were very positive*
  - c60% of reviewed cases were high risk
  - c25% of the high risk referrals were unique to Trinity
  - The model identified a good mix of fraud types
- *The model now implemented with fraud operations*
  - Scores are refreshed daily
  - Trinity scores are crossed with a fraud propensity model to drive referrals



Strategy
Matrix of Trinity & Fraud Score

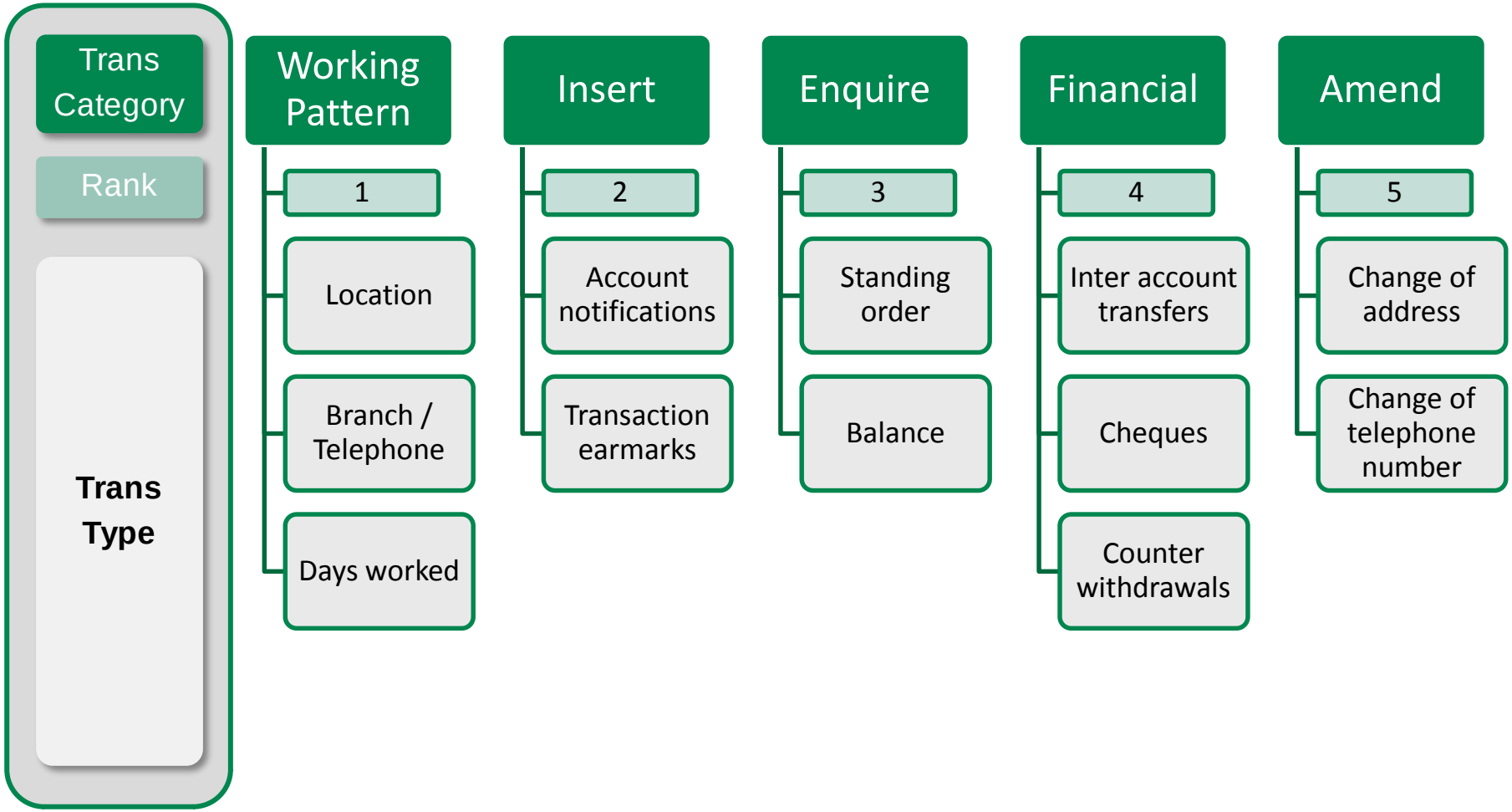


INSIDER FRAUD

DETECTING INSIDER FRAUD
USING COLLEAGUE
INTERACTIONS

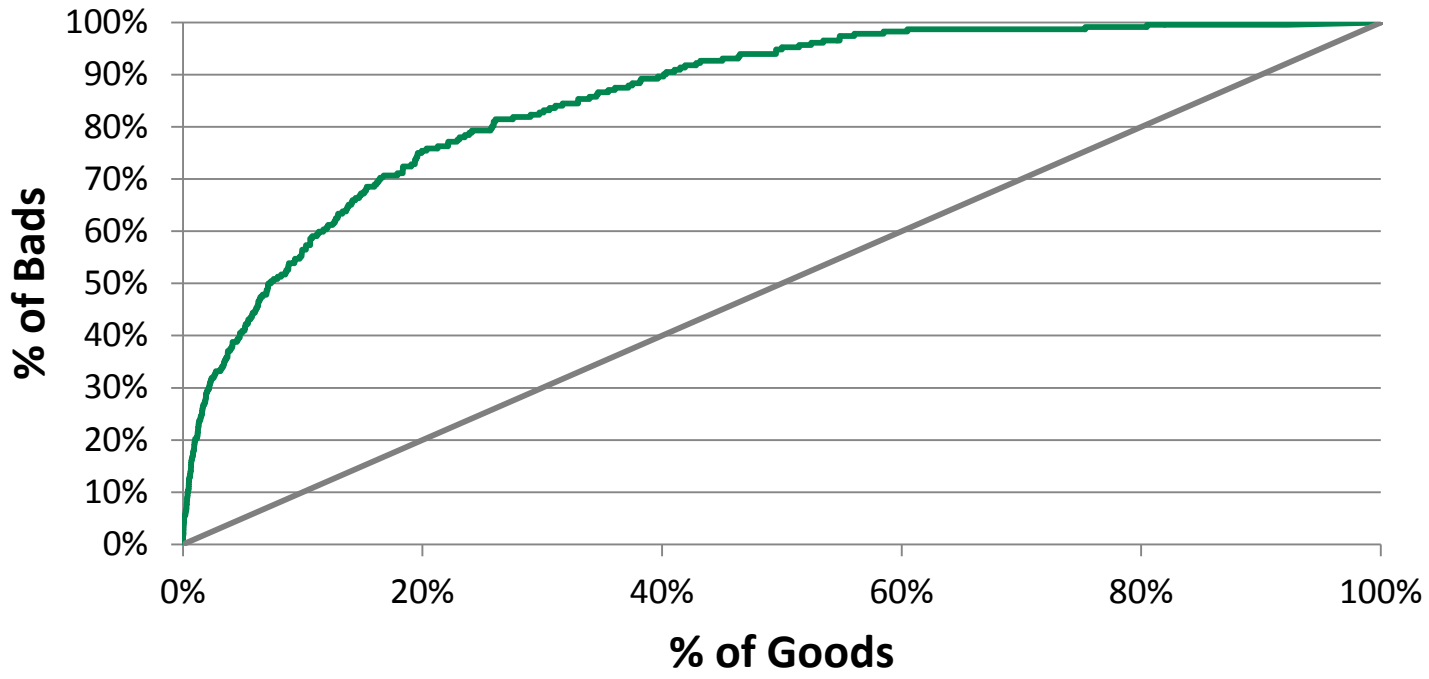
MODEL DRIVERS

The model uses hundreds of event types as predictors, these are driven by colleagues and related to system log on identifiers



MODEL PERFORMANCE

Model performance on the validation sample is strong



- Performance is shown for the development validation data set
- Discrimination is very high: Gini c**70%**
- The model discriminates particularly well at the top of the distribution
- Initial use case will be to exclude all staff members from investigation where the model shows sufficiently low risk

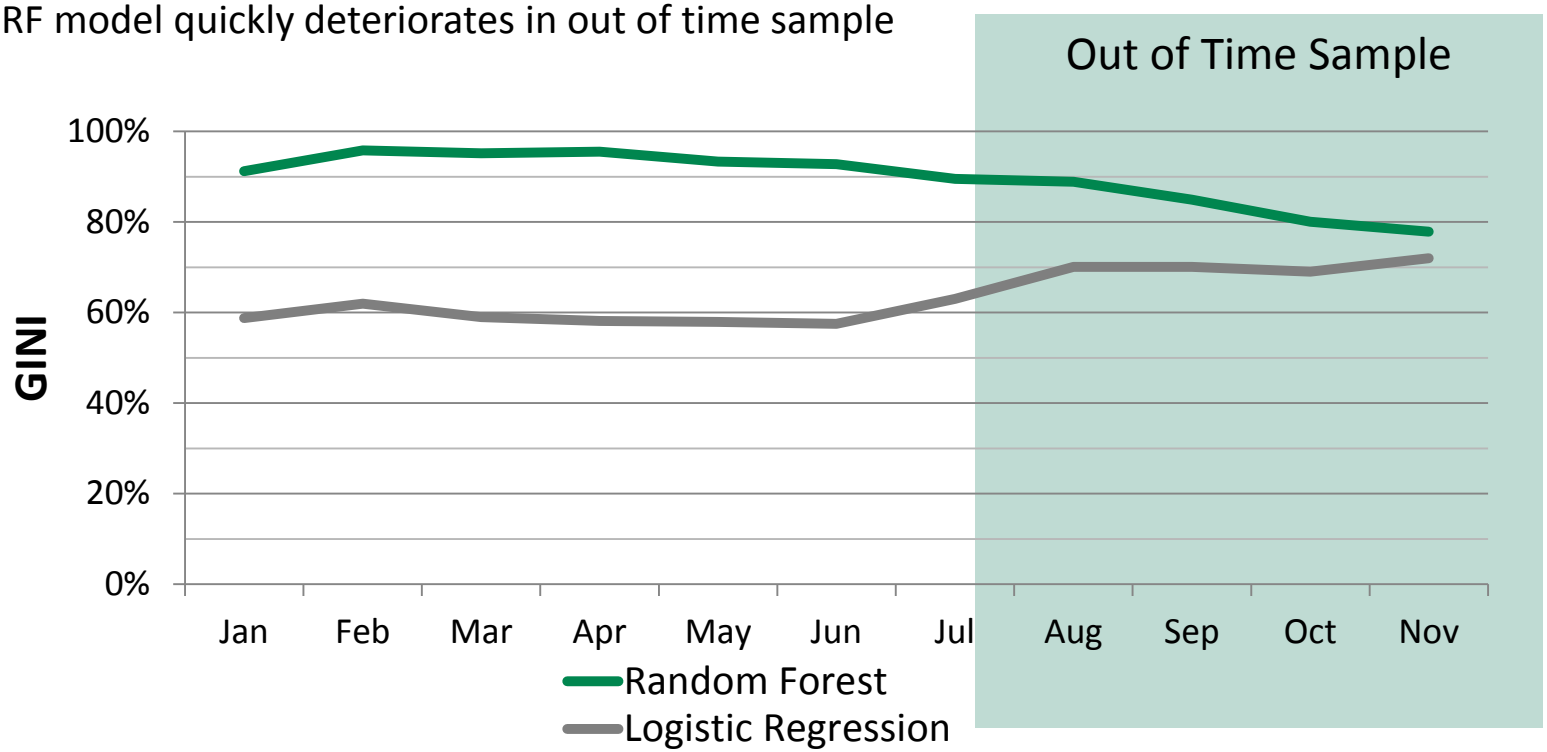
MORTGAGE BROKER RISK

ASSESSING THE RISK OF MORTGAGE BROKERS USING APPLICATION MIX

MORTGAGE BROKER RISK

UNABLE TO VALIDATE RESULTS OUT OF TIME

- Model developed to assess risk posed by Mortgage Broker Panel
- Over 55,000 characteristics assessed relating to application mix
- Random Forest and logistic regression models developed
- RF model quickly deteriorates in out of time sample



- Desire to investigate further. Is the algorithm overfitting to the development sample?
- Lack of resource to investigate at this time

CONCLUSION

CONCLUSION

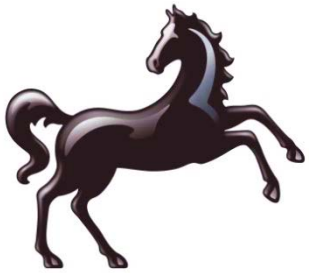
Random Forests can add value to fraud prevention

Positive Results

- Very high discrimination
- Relatively quick to build
- Many predictors used to assess risk
- Hard for fraudsters to game

Ongoing Challenges

- Validation issues. Are the models more likely to over fit to the development period?
- Implementation challenges. Currently no decisioning platforms for deployment
- Monitoring. What to monitor on monthly / quarterly basis?
- Desire to develop best practice for development and ongoing monitoring and maintenance



LLOYDS BANK