Credit scoring

1. Credit scoring for consumer credit
2. Measuring performance using the lift chart

1. Credit scoring

- Want to predict probability of loan non-repayment, and hence approve/reject credit request 1000 records containing information about previous loans (300 default)
- Sex, Marital Status, Age, Years residence at current address, House owner, Telephone available, . . . , Bank account holder, History of previous repayments, Debts, Employment type, . . . , Loan amount, purpose, loan period
- 20 input variables in total

Building the model set

- Data for mining should be in a single table
- Each row should correspond to a single instance - a “unit of action”
  - So: need to understand how the results of the mining effort will be employed
  - Here the unit of action is a “loan”
  - In other cases it could be: customer, insurance policy, insurance claim, basket, household, PC, inventory item

Oversampling

- Suppose in reality we had loans with 1% non-repayment. A model can easily make a 99% accurate prediction
  - i.e., “every loan does not default”
- *Oversampling* is the process of taking more of the rarer outcomes
  - Aim for 20-30% “density” of the rarer outcome

Transforming the input data

- Some variables are flattened
  - e.g., Account is transformed into 2 binary variables: good account and bad account

good_acc	**bad_acc**	**account details**
1	0	balance > 200
0	1	balance < 0
0	0	balance in 0 ..200
0	0	no account

Transforming the input data

- Others are transformed to binary via: value > median and value ≤ median
- Data matrix

Applicant	Default	X1	X2	X3	…
1	0	1	1	1	…
2	1	1	1	1	…
3	1	0	0	0	…
…	…	…	…	…	…

The median

A
S
I
D
E

The median of a set of values

$$x_1 < x_2 < x_3 < x_4 < . . . < x_n$$

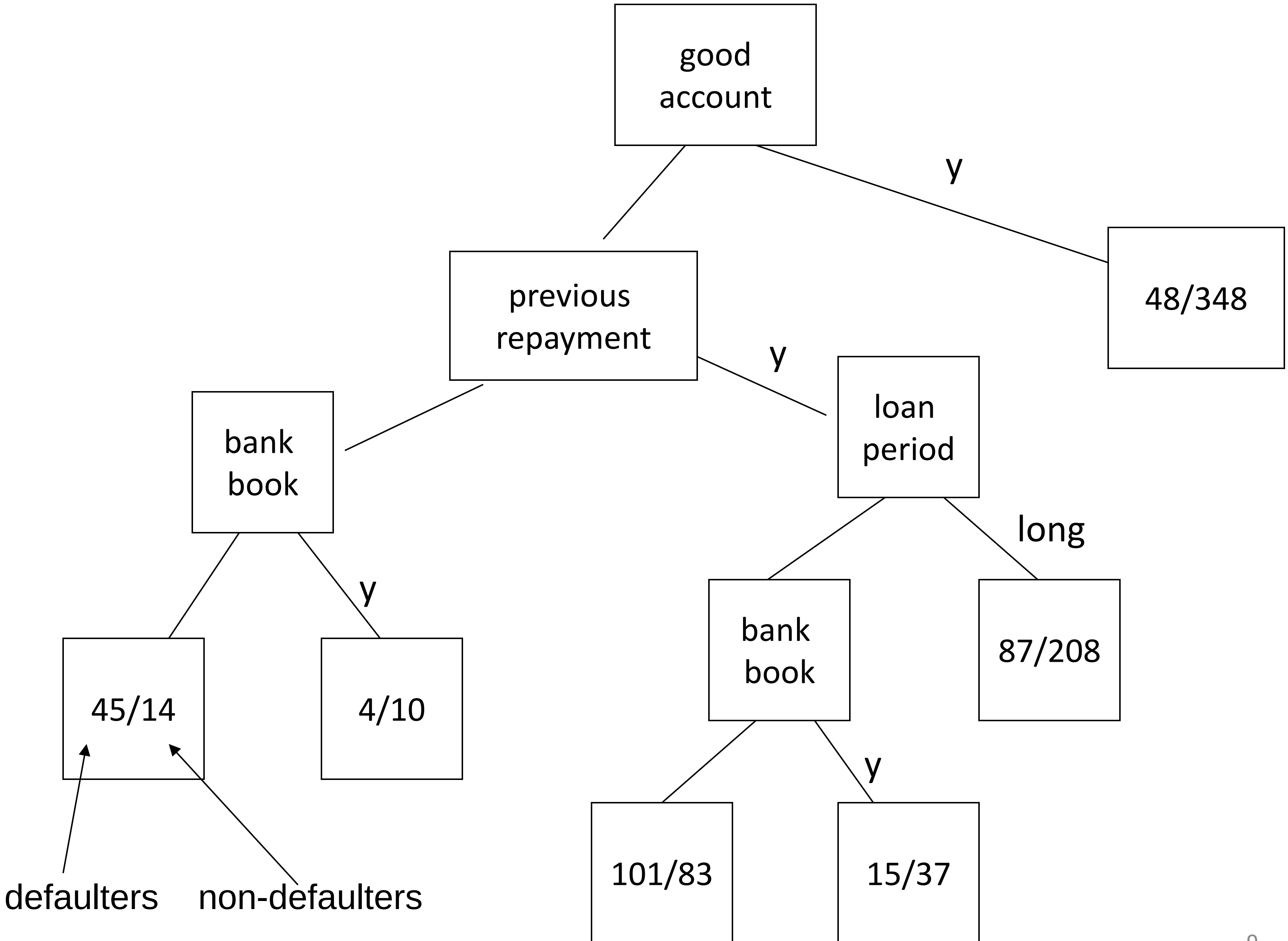
is the “middle value”

If n is odd, the median is x_k where $k = (n+1)/2$

If n is even, the median is the average of the two middle values x_k (where $k=n/2$) and x_{k+1}

The resulting decision tree

- 6 leaf nodes based upon 4 variables
  - good account
  - bank book
  - previous repayment history
  - loan period (long/short)


good account
y
48/348
previous repayment
y
loan period
long
87/208
bank book
y
45/14
4/10
bank book
y
101/83
15/37
defaulters
non-defaulters

The resulting leaf nodes

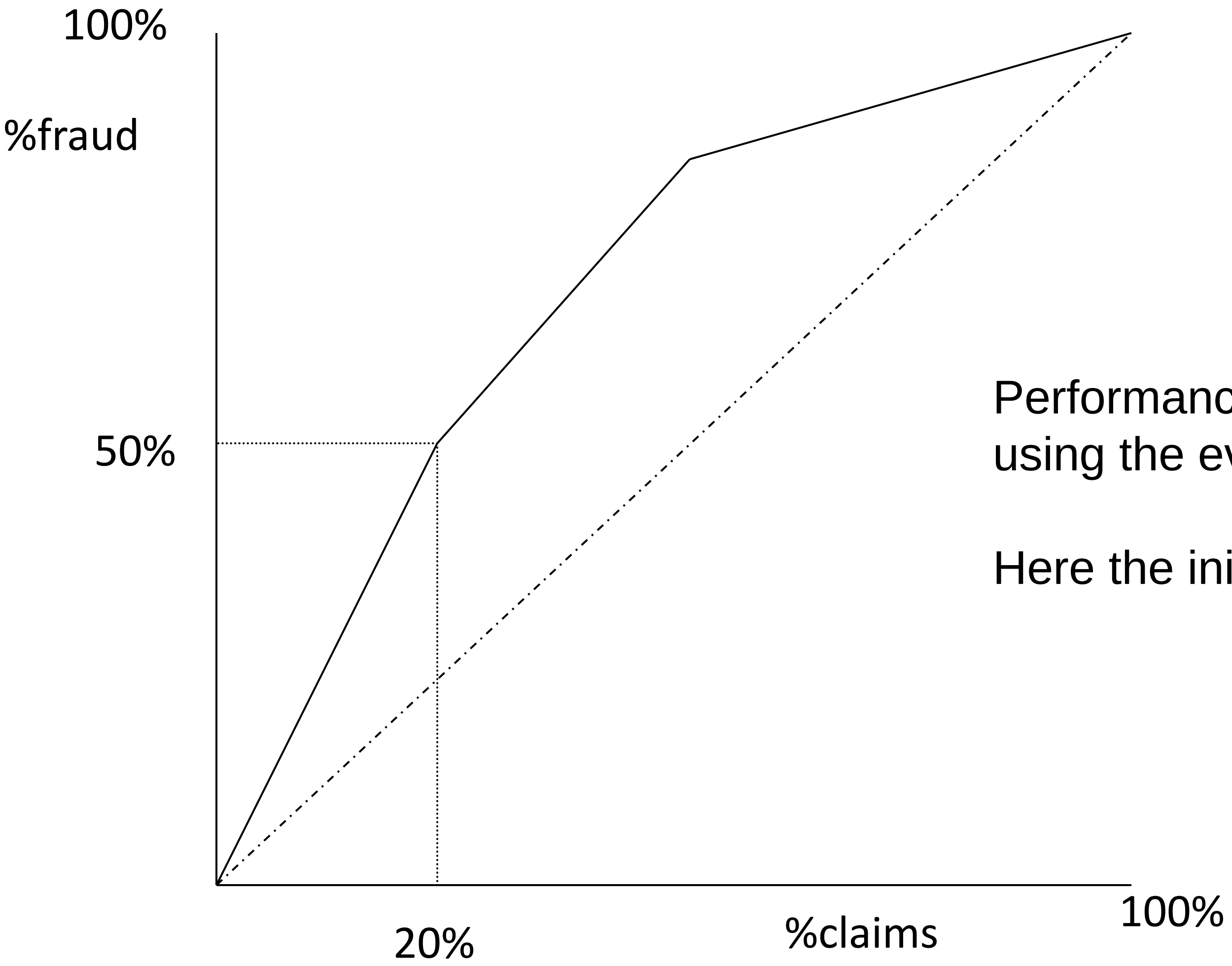
Node	Number of data items	Defaulters	%Defaulters
45/14	59	45	76%
4/10	14	4	29%
101/83	184	101	55%
15/37	52	15	29%
87/208	295	87	29%
48/348	396	48	12%

We then need to decide which of these nodes is acceptable for loan approval. Since there is little to choose between those labelled with 29% we might either include them all or exclude them all. Suppose that we decide to include all of these for approval, then:

The resulting rules

- if good_account = n and previous_repayments = y and loan_period = short and bank_book = n then not approved
  - This is the branch to the leaf 101/83
- if good_account = n and previous_repayments = n and bank_book = n then not approved
  - This is the branch to the leaf 45/14
- else approved

2. Measuring performance



A lift chart

Performance should be assessed using the evaluation set

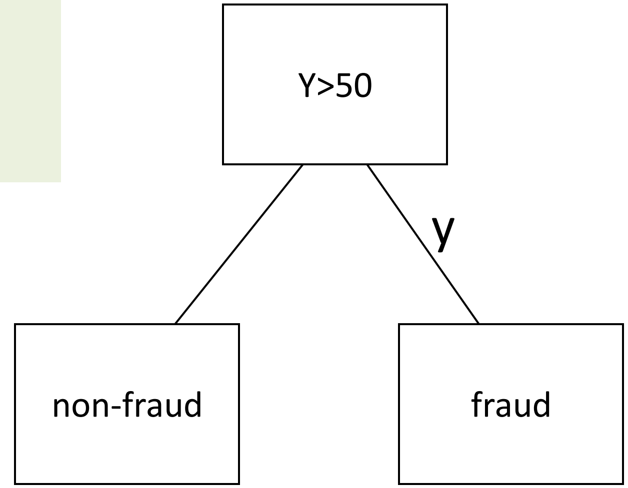
Here the initial lift = 50/20

Interpreting the lift chart

- Suppose that we have 10000 claims of which 3% (ie 300) are fraudulent
- Using the previous lift chart the top 20% of the claims (20% of 10000 = 2000) contain half the fraud (ie 150)
- Given a claim in the top 20%, probability it is fraudulent =150/2000=7.5%
- Lift = gradient = 5/2 = the improvement in density of targetted outcome = 7.5/3

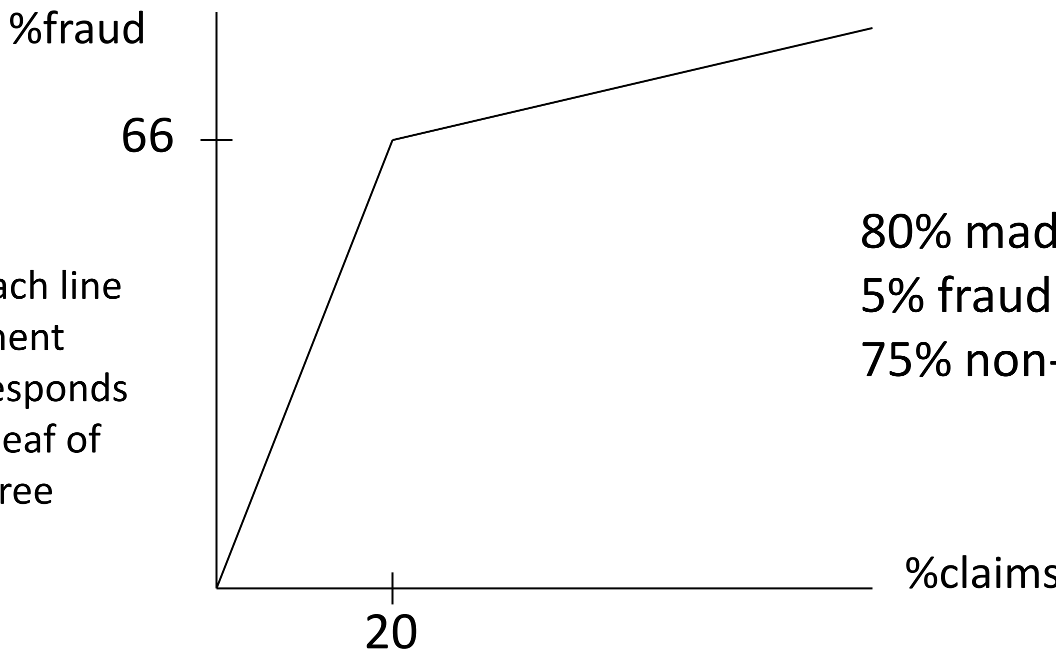
Lift charts for decision trees

Suppose that in the data (evaluation) set we have 15% fraud altogether. Suppose 20% of the individuals have Y>50, and that 50% of this subset are fraudulent



80% made up of
5% fraud
75% non-fraud

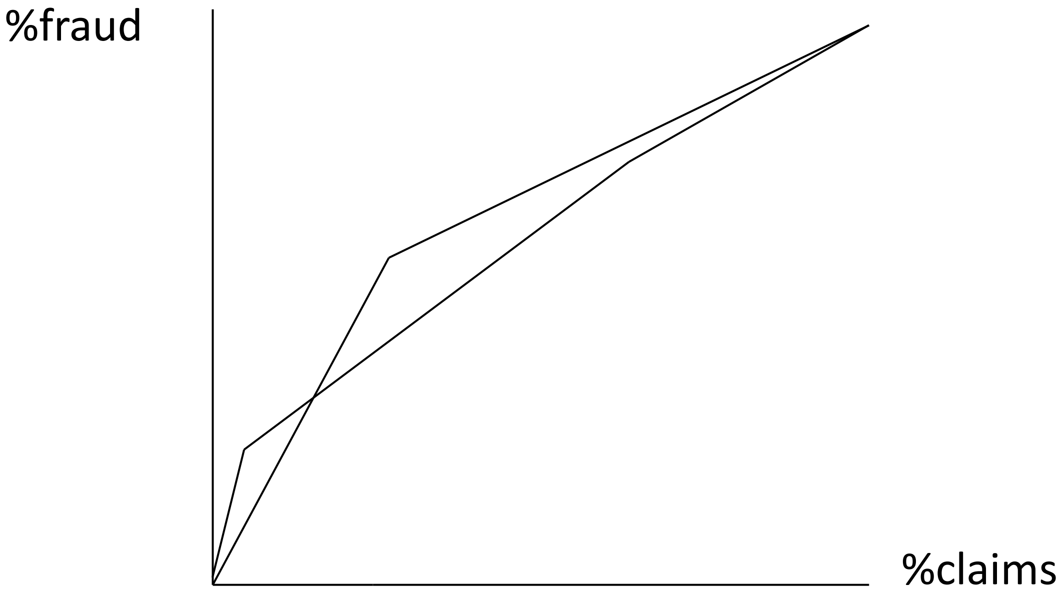
20% made up of
10% fraud
10% non-fraud



So each line segment corresponds to a leaf of the tree

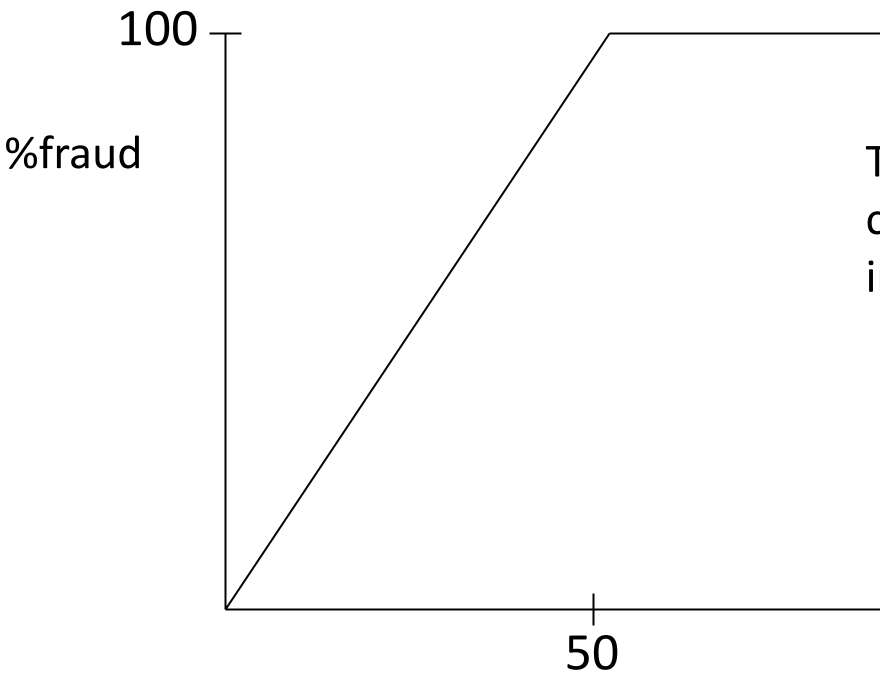
Lift charts for comparing models

Which model is best?



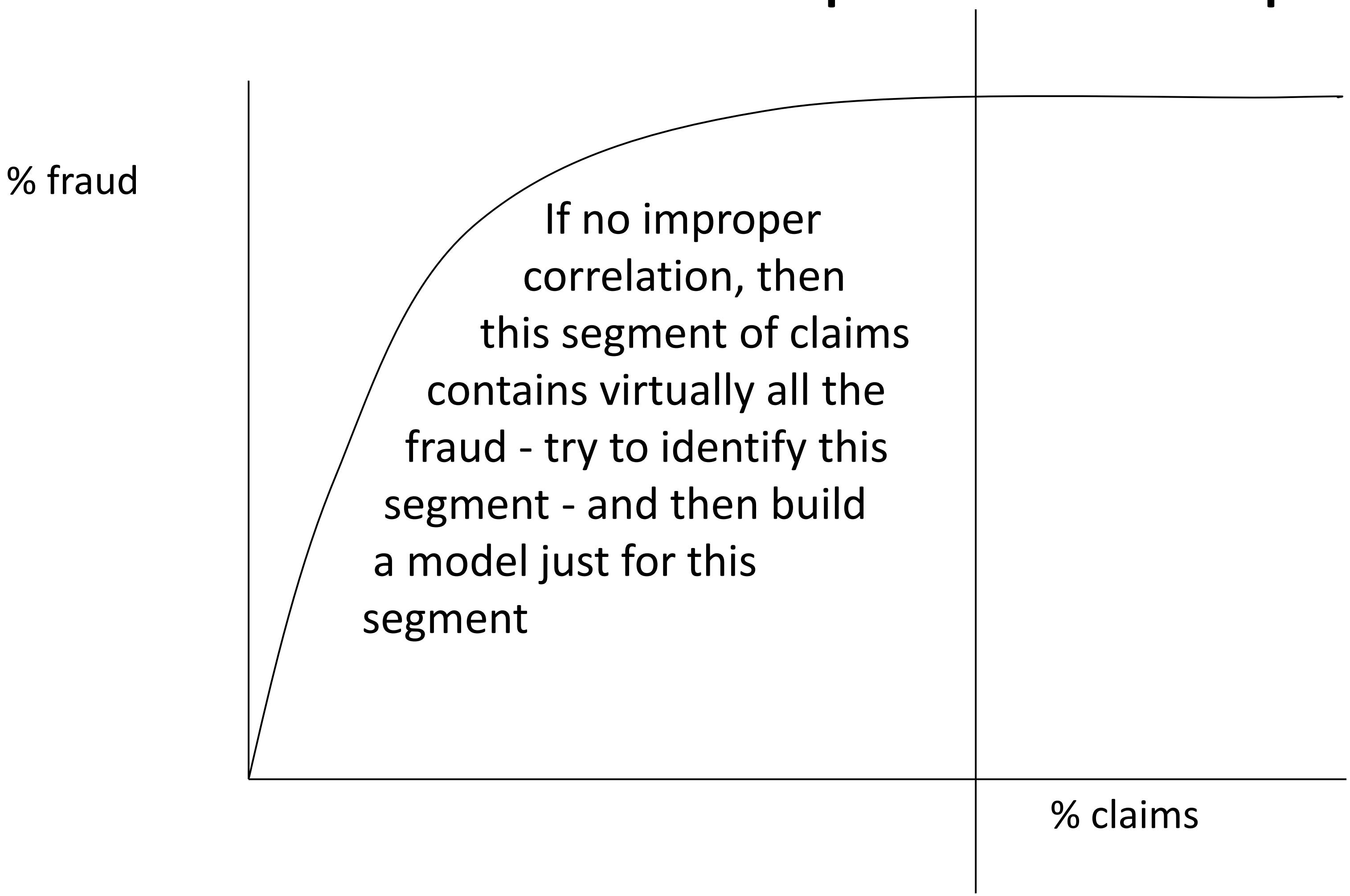
There are theoretical limits to lift

Suppose that in the data (evaluation) set we have 50% fraud altogether, and that our predictions are perfect! The lift is still only 2.



The limits on lift are a consequence of the initial high density

A flat line may indicate an improper correlation between inputs and outputs



The lift chart for the credit scoring application