



13th - 17th May 2019 – Kuala Lumpur

Fraud Model Development and Deployment in SAS FM

Session 4: Data Preparation

Data Preparation

Topics

- Review of common data sources
- Entities and house-holding
- Fraud matching and tagging
- Sampling methodology
- Basic statistics and data validation



Data Preparation

Data Sources

Data Preparation

Data Sources for Historical Data

- For first time models, transactional data need to be provided by the bank
- Typically data sources are partitioned along the following categories:
 - Monetary transactional data
 - Authorizations, payments, deposits
 - Typically originates from the primary channel on focus
 - Non-monetary transactional data
 - Demographic information changes, account maintenance activities, session data
 - Can originate from various channels
 - Master-files
 - Slow changing non-monetary transaction that are not event based
 - Periodic snapshots of customer data, bureau data etc.
 - Fraud data
 - Typically maintained as a separate source
 - Can come from reporting sources such as SAFE / TC40
 - Will need to be matched to one or more transactional sources

Data Preparation

Historical Data Mapping

- Production data comes from production sources
 - Core banking systems
 - Real-time payment systems
 - Batch processes
- When mapping historical data to SAS messages, extreme diligence should be exercised to ensure that the resulting data will very closely resemble data that will be generated in the production environment
 - Any deviations should be manually treated within the model code
 - Hence it is ideal to work on historical and production data mapping as concurrently as possible
 - Generally not an issue when historical data comes from consortium processes



Data Preparation

Entities and House Holding

Data Preparation

Entity

- There is always one or many entities involved in a transaction
- E.g.
 - Account number
 - Customer ID
 - Card number
 - Beneficiary account number
 - Beneficiary bank
 - Merchant ID
 - Terminal ID
 - IP address
 - Shipping address

Data Preparation

Entities

- Models are always built around one or more entities
 - Behavior is tied to entities
 - Entities drive signatures
 - Theoretically possible to built a model without considering any entity
 - That means any behavioral information cannot be utilized by the model
 - Based purely based on population risk factors
- Choice of entities around which the model is built determines subsequent modeling steps:
 - Fraud tagging
 - Sampling
 - Features engineering

Data Preparation

Entity Relationships

- When more than one entity is being tracked by signatures, careful attention should be paid to the relationship between them.
- Entities can have different relationships between them:
 - 1 – to – 1
 - 1 – to – Many
 - Many – to – Many

Data Preparation

Entity Relationships

- 1 – to – 1
 - E.g. card number and account number in some credit card portfolios
 - Redundant to track the same activity on both entities if both are specified in the transaction
 - There may be transactions where only one entity is specified; it may be required to have those activities tracked by the appropriate entity signature
 - E.g. card reissue event may be reported only with the card number

Data Preparation

Entity Relationships

- 1 – to – Many
 - E.g. card / account numbers in debit card portfolio / account and user ID
 - The larger entity will be able to capture all activities across the smaller entities associated with it
 - Can just track the smaller entity for each event within the signature itself
 - However having separate signatures may be able to provide a more granular view
 - Also some events may be reported only at the smaller entity level

Data Preparation

Entity Relationships

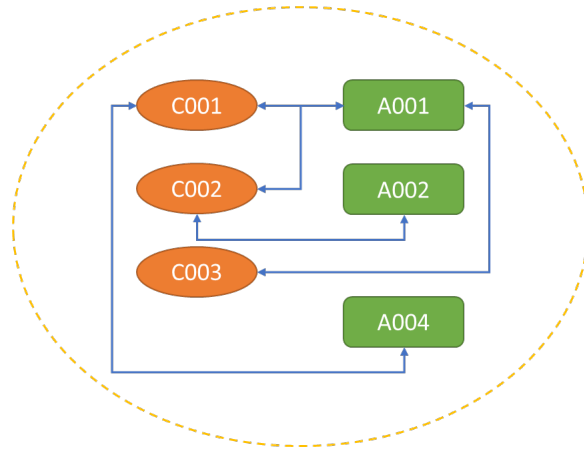
- Many – to – Many
 - E.g. card number and account number in most card portfolios / customer IDs – user IDs
 - Necessary to track the activities on both entities

Data Preparation

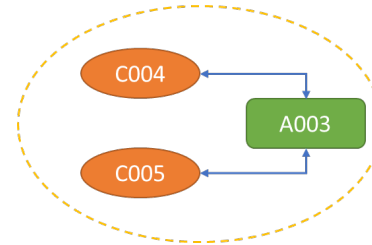
Householding

- Related entities in a many – to – many relationship form a “cluster” – termed as a household
- Entities within a household are connected components

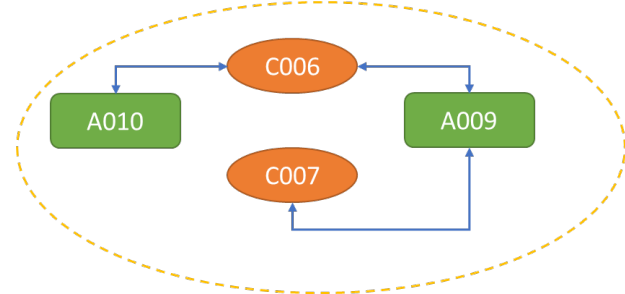
Card	Account
C001	A001
C002	A001
C002	A002
C005	A003
C003	A001
C004	A003
C006	A010
C007	A009
C001	A004
C006	A009



Household 1



Household 2



Household 3

Data Preparation

Householding

- In a 1 – to – many relationship scenario, the household degenerates into just the larger entity
 - For e.g. if multiple accounts are linked to a single user ID, the user ID effectively becomes the household.
- The notion of household does not apply in a 1 – to – 1 relationship

Data Preparation


Householding with PROC OPTNET

```
data entities;
  /* PROC OPTNET requires that the input dataset specify the
     linkages via a 'from' and 'to' variables */
  from = "C001"; to = "A001"; output;
  from = "C002"; to = "A001"; output;
  from = "C002"; to = "A002"; output;
  from = "C005"; to = "A003"; output;
  from = "C003"; to = "A001"; output;
  from = "C004"; to = "A003"; output;
  from = "C006"; to = "A010"; output;
  from = "C007"; to = "A009"; output;
  from = "C001"; to = "A004"; output;
  from = "C006"; to = "A009"; output;
run;

proc optnet data_links = entities out_nodes = households;
  concomp;
run;

proc sort = households;
  by concomp;
run;

proc print = households;
run;
```



Obs	node	concomp
1	C001	1
2	A001	1
3	C002	1
4	A002	1
5	C003	1
6	A004	1
7	C005	2
8	A003	2
9	C004	2
10	C006	3
11	A010	3
12	C007	3
13	A009	3

Data Preparation

Importance of Householding

- When modeling using multiple signatures in 1 – to – Many or Many – to – Many cases, it is essential that the appropriate households are generated.
- If an entity from a household is sampled for modeling, then all other entities that belong to the household should also be sampled in.
 - Failure to do so will lead to ‘gaps’ in signature when emulating the signature during model development
 - This will lead to a model trained on signature states that will not be observed in production
 - Will have adverse impact on model performance

Data Preparation

Importance of Householding

- Consider a model that uses a card and account signatures
- Suppose householding was not applied when sampling data

During
Model
Development

Card	Account	Merchant	Amount	POS
C001	A001	MARATHON #1567	77.49	02
C002	A002	BANANA INTERNATIONAL PLA	102.08	90
C001	A001	OLD NAVY-ON DRCT ONLINE	40.54	01
C002	A001	CHEVRON UTC	45.22	90
C001	A002	ANTHROPOLOGIE #569	31.19	90



Merchant	Amount	POS
MARATHON #1567	77.49	02
OLD NAVY-ON DRCT ONLINE	40.54	01

Merchant	Amount	POS

=

≠

In Production

Card	Account	Merchant	Amount	POS
C001	A001	MARATHON #1567	77.49	02
C002	A001	BANANA INTERNATIONAL PLA	102.08	90
C001	A002	OLD NAVY-ON DRCT ONLINE	40.54	01
C002	A001	CHEVRON UTC	45.22	90
C001	A002	ANTHROPOLOGIE #569	31.19	90



Merchant	Amount	POS
MARATHON #1567	77.49	02
OLD NAVY-ON DRCT ONLINE	40.54	01

Merchant	Amount	POS
BANANA INTERNATIONAL PLA	102.08	90

Card Signature

Account Signature

Data Preparation

Householding

- Householding is not suitable for certain entity combinations
- For e.g. consider a model that employs a card and terminal signature
 - Since a terminal can interact with a large number of cards, this will result in a small and meaningless number of households
 - In such cases, we rely on other methods to emulate the signature without gaps
 - Will be covered in a later session
- Other examples for such cases are:
 - Account and beneficiary account signatures
 - Certain large beneficiary accounts such as utility accounts impact the households
 - User ID and IP address signatures
 - Some IP addresses can be problematic



Data Preparation

Fraud Tagging

Data Preparation

Fraud Matching

- Fraud transactions may be provided as a separate data source
 - Needs to be 'matched' to actual transaction within the transactional dataset
- May have varying levels of differences compared to the transactional data source:
 - Some attributes may not be available
 - Format differences
 - E.g. 12 hr vs 24 hr notation of time,
 - Value differences
 - E.g. amounts be in different currencies, authorization vs posted amounts, merchant location
- Typically needs a fuzzy process to perform this matching
 - Won't get into the details of it here

Data Preparation

Fraud Matching

Date / Time		Amount	MCC	Merchant	Zip Code	POS
20140301	09:27:10	77.49	5542	MARATHON #1567	92137	02
20140302	13:23:01	102.08	5713	BANANA INTERNATIONAL PLA	92122	90
20140303	12:51:35	40.54	5713	OLD NAVY-ON DRCT ONLINE	94105	01
20140304	18:12:55	45.22	5542	CHEVRON UTC	92122	90
20140304	18:46:20	31.19	5651	ANTHROPOLOGIE #569	92122	90
20140304	19:21:43	42.12	5712	THE LAND OF NOD 158	92122	90
20140304	19:38:25	2.75	5814	CHAMPAGNE BAKERY 4207	92122	05
20140306	15:20:07	78.56	5713	AMAZON RETAIL	92122	81
20140307	08:38:45	1.00	5542	MARATHON PETRO041350	92064	02



4/3/2014	45.22	5542	UTC CHEVRON GAS	92122
----------	-------	------	-----------------	-------

Data Preparation

Fraud Matching

- The outcome of fraud matching process is to have as many transactions in the fraud source matched as possible (> 95%).
 - Low match rates can be indicative of data gaps or underlying data issues
 - Should be thoroughly investigated and fixed
 - Unmatched transactions have a dual negative effect
 - Reduces the frauds to learn from
 - Introduces noise in the non-frauds
- Matched transactions are considered the actual fraud transactions in subsequent steps.

Data Preparation

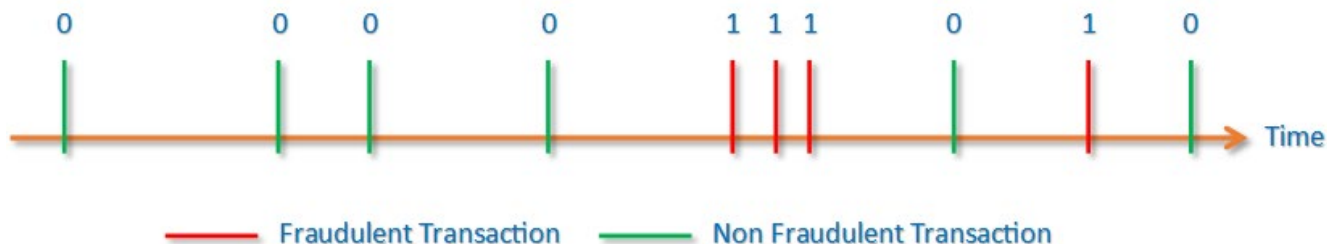
Fraud Tagging

- Fraud tagging is the step in which model targets are defined.
- Depending on the fraud problem, tags can take various definitions
 - Fraudulent state of the transaction
 - Only fraud transactions are the target
 - Fraudulent state of the entity
 - All fraud transactions after the compromise are the target

Data Preparation

Transaction Based Tagging

- The actual fraud events are tagged as '1'
- All other events are tagged as '0'



Data Preparation

State Based Tagging

- In state based tagging, the transactions for a fraud entity are considered to belong to one of three states:
 - Pre –fraud: all transactions before the first fraud
 - Given a tag value of '2'
 - Post – fraud: all transactions on and after the fraud is detected / blocked
 - Given a tag value of '3'
 - Fraud window: any transactions between the pre fraud and post fraud period when the entity is in an 'active state of fraud'
 - Given a tag value of '1'
- Transactions from non-fraudulent entities are always in a state of non-fraud
 - Tagged as '0'

Data Preparation

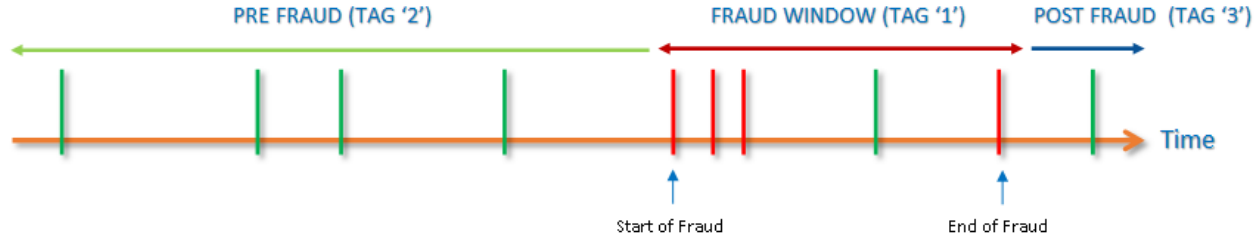
State Based Tagging

- First fraud is almost always the first fraud transaction
- There are three ways to determine when the block goes into effect:
 1. Based on the last approved transaction
 2. Based on the last approved fraud transaction
 3. Based on an explicitly provided data on when the entity was blocked
- No standard guideline to pick which method to use
 - In practice both of them will yield good results
 - Will have an effect on the replacement window (will be discussed shortly) and performance evaluations

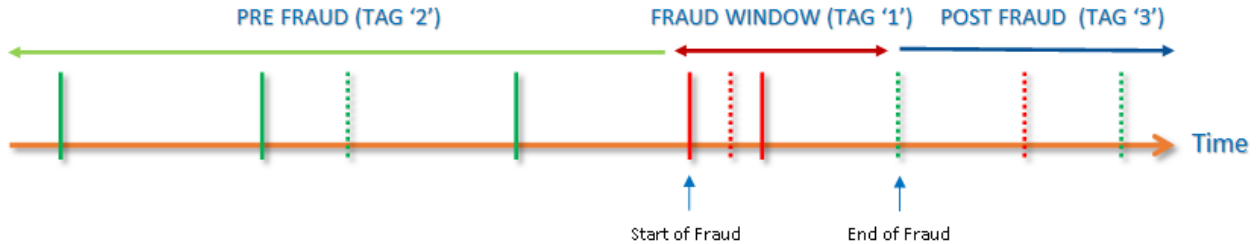
Data Preparation

State Based Tagging

Based on last fraud transaction



Based on last approved transaction

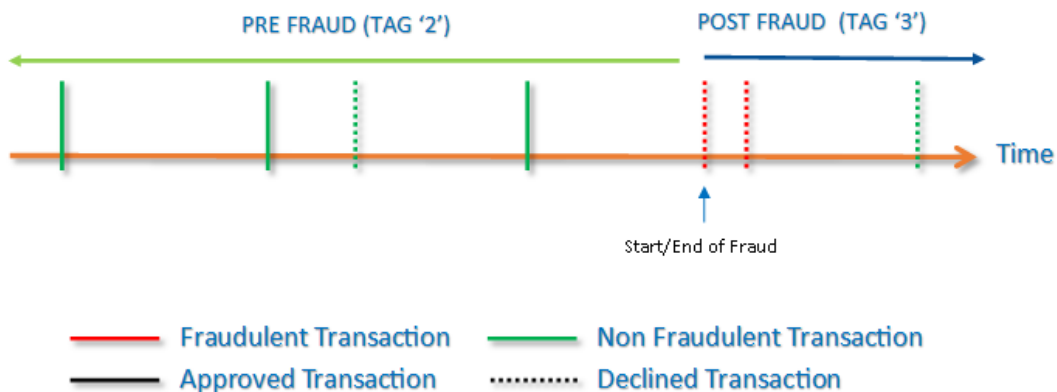


— Fraudulent Transaction — Non Fraudulent Transaction
— Approved Transaction Declined Transaction

Data Preparation

State Based Tagging – No Loss Cases

- There may be cases where the fraud was detected and the entity blocked on the very first transaction using the current fraud system
- Such cases may not have an active fraud window :



Data Preparation

State Based Tagging – Tags to Targets

- The target in state based tagging is whether the entity is in a state of fraud or not
 - Not if a transaction is fraud or not
 - There is a fundamental difference between these two
- Therefore tag 1s will represent the fraud targets in the modeling data
- We typically exclude tag 2s from the final modeling dataset
 - They are just like tag 0s and we have plenty of them
 - There may also be ambiguities in fraud boundaries
 - Retained till signatures are appended
- Some tag 3s are converted to tag 1s for multiple reasons.

Data Preparation

Transaction vs. State Based Tagging

Transaction Based	State Based
Suitable for cases where the entity cannot be associated with a fraud state. E.g. deposit fraud	Suitable for cases where the tagged entity goes into a state of fraud. E.g. card fraud
Model trained to detect if a transaction is fraud or not	Model trained to detect if an entity is in a state of fraud or not
No replacement window	Need to consider replacement window
Not tags to exclude	Tag 2s and most 3s are excluded
Models are generally trained with signature contamination	Models are trained without signature contamination

Data Preparation


Which Entity to Tag?

- Though there may be more than one type of entity in the household associated with the model, tagging is done at a single entity level
 - Fraud is almost always reported at a single entity level
- For state based tagging, this eventually means the model score predicts if this particular entity is in a state of fraud or not
- Most of the time, this entity will be very evident from the onset based on business requirements and / or nature of the problem and data

Data Preparation

More on Tagging

- The following cases can be observed at an individual entity level when combined with households:
 - Good household – good entity
 - These entities are all tagged as 0's as described earlier
 - Fraud household – fraud entity
 - These entities are all tagged based on state or transaction as described earlier
 - Fraud household – good entity
 - These entities are typically tagged with a value of 4
 - Mostly excluded from the modeling data after appending the signature state to the transaction
 - Needs careful treatment if including in the modeling data



Data Preparation

Sampling Methodology

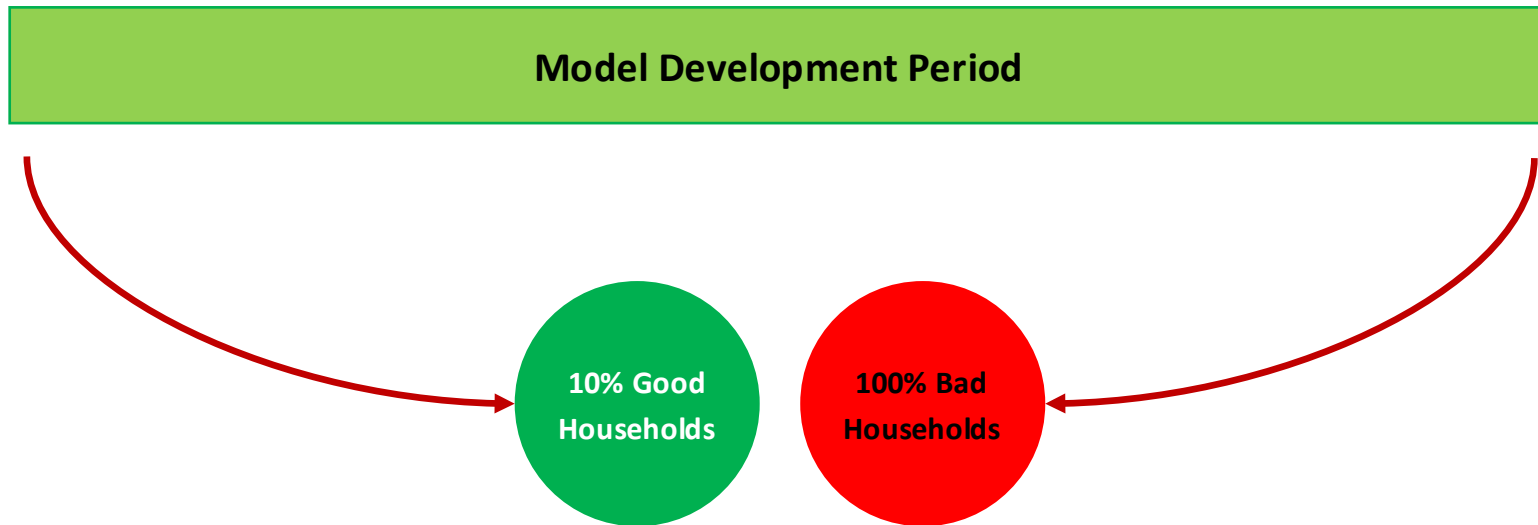
Data Preparation

Sampling Methodology

- Since fraud is a very rare event, we typically keep all the frauds in our modeling dataset.
 - i.e. 100% of all fraud households are retained in the modeling dataset
- Non fraud households are typically sampled down
 - Typically 5 – 10 % of the good households are retained
 - Function of imbalance and associated volume of data
- Upsampling very rarely used
 - Overfitting issues
 - Upsampling to match the goods will result in a massive dataset

Data Preparation

Sampling Methodology





Data Preparation

Partitioning Final Datasets

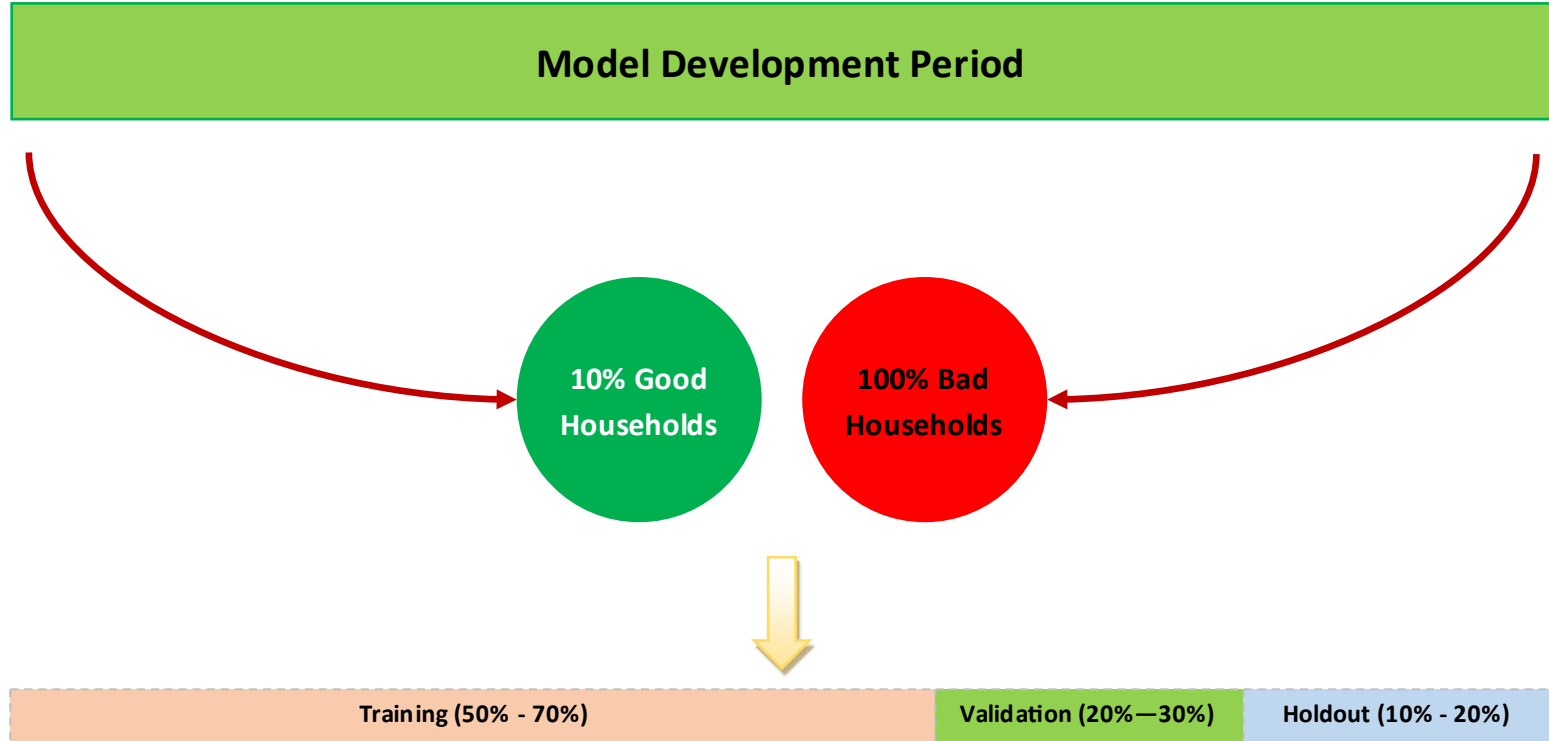
Data Preparation

Partitioning Final Datasets

- The resulting sample dataset is broken down into three parts:
 - Training
 - Validation / test
 - Holdout
- Holdout dataset is a subset of data set aside to be used only for final in-period performance reporting
- In very low fraud volume problems, holdout can be skipped
 - An out-of-time dataset can be used for performance reporting

Data Preparation

Partitioning Final Datasets



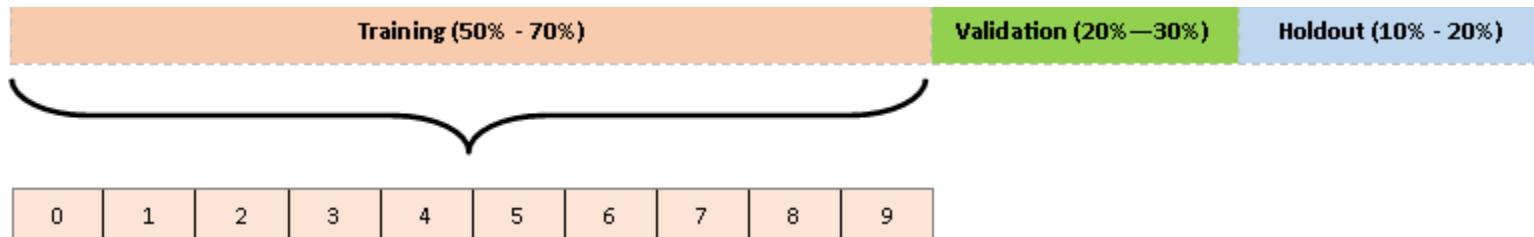
Data Preparation

Further Partitioning Final Datasets

- In order to parallelize various steps, the partitions can be further broken down into multiple chunks
- We can assign a random digit to each household
 - Each chunk can hold a certain range of these random digits
 - For e.g. part 0 of the training dataset can contain households with random digits 0 – 99, part 1 of the training dataset can contain households with random digits 100 – 199 etc.
 - Data for a given household should reside within the same chunk
- Various steps can be run for each chunk in parallel
 - Signature enrichment, feature generation
- Results from the parallel runs can be aggregated for other steps
 - Feature selection, model training etc.

Data Preparation

Further Partitioning Final Datasets





Data Preparation

Basic Statistics and Data Validation

Data Preparation

Basic Statistics and Data Validation

- Once we have the modeling datasets, a set of basic statistics can be generated based on a portion of the training dataset for data validation
- A simple **PROC FREQ** and **PROC MEAN** applied on fields used for modeling, broken down by each month can reveal data issues
 - Distributions should be stable month – over –month
 - Any deviations should be explainable through expected seasonal variations

Data Preparation

Sample Basic Statistics



Basic Statistics