CLUSTERING AND IMPLEMENTATION

Dr. Aric LaBarr
Institute for Advanced Analytics

Course Layout

Data Preparation
- Transactional Data
- Recency vs. Frequency
- Network Features

Anomaly Models
- Univariate Analysis
- Clustering
- Isolation Forests
- CADE

Fraud Supervised Models
- SMOTE
- Models
- Labeled vs. Unlabeled Bias
- Not Fraud Model
- Evaluation

Clusters of Not Goods
- Cluster Analysis
- Social Network Analysis

Implement
- Investigators
- Traffic Light Indicators
- Backtesting

Fraud Maturity

Components	New / Young	Emerging SIU	Fraud Scoring	Holistic Solution
Simple Rules	Yes	Yes	Yes	Yes
Unlabeled Data	Yes / No	Yes / No	Yes	Yes
Labeled Fraud Cases	No	Yes	Yes	Yes
Anomaly Models	No	Yes / No	Yes	Yes
Supervised Models	No	No	Yes	Yes
Non-Fraud Models	No	No	No	Yes
Clusters of not Good	No	No	No	Yes

CLUSTERS OF NOT GOODS

Fraud Model, Not-Fraud Model, …

- After identifying both the fraud and not-fraud models from the known data, turn attention to **unknown** data.
- Trying to find the unique instances of observations that aren't like previous fraud **and** not like previous not-fraud.

Unknown **Scored** Observations

- Possibly too many to investigate, so how do I prioritize the ones I need.
- Instead of just giving highest scoring observations, sometimes we take same approach as when we didn't have data:
  1. Anomaly models
  2. Clustering

Unknown **Scored** Observations

- Find the collections of **scored** observations that might represent **new** groups of fraud.
- Then same process with SME’s as before:
  1. Subject matter experts will look through the suspected anomalies (clusters) for cases that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.
  3. Treat these suspected fraud groups as if they had committed fraud and other groups as if they have not.
  4. Ideally, have subject matter experts also identify small set of legitimate claims in non-anomaly data.

Unknown **Scored** Observations

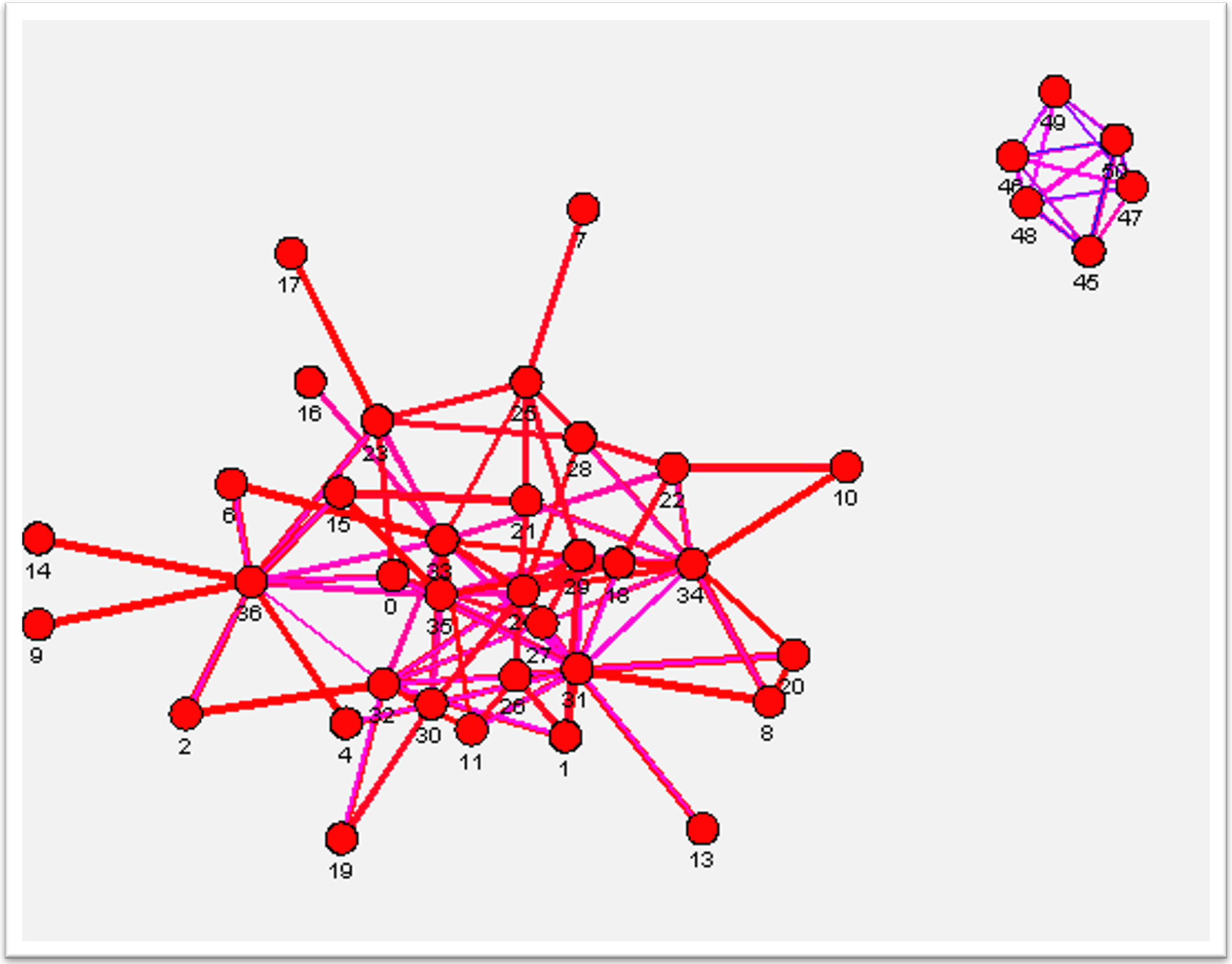
- One of 2 paths:
  1. **IDEALLY**, investigators trust your process and investigate new types of fraud based solely on the SME recommendations.
  2. **MIGHT** have to put these tagged “possible new fraud” claims into the modeling process and let the model results tell the investigators to act.

NETWORK ANALYSIS FOR UNKNOWNS

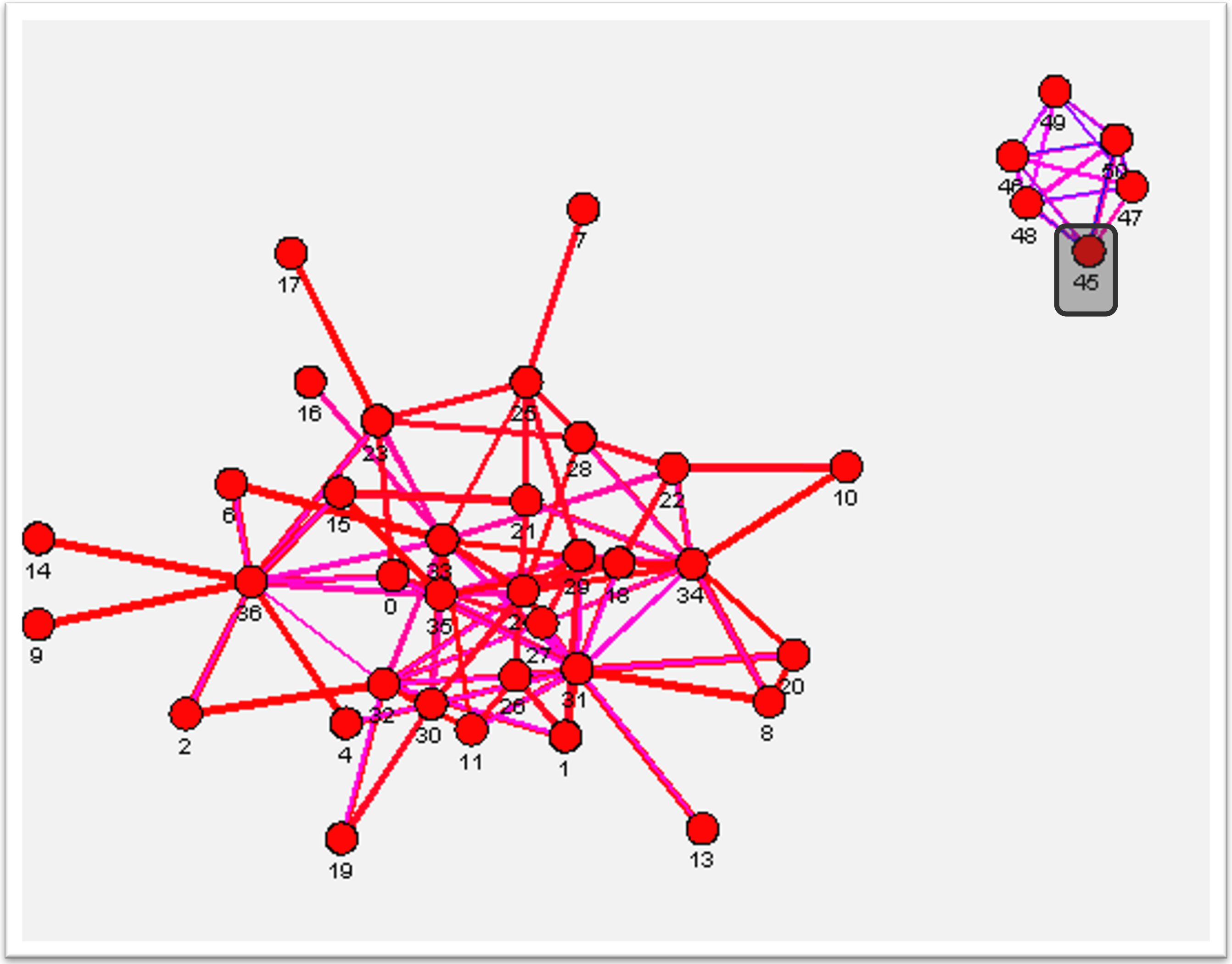
Subgroups

- Social networks typically contain dense pockets of individuals.
- These dense pockets are sometimes called **subgroups**.
- If a subgroup is completely separated from the rest of the network, then it is a **cohesive subgroup**.
- Homophily: “Birds of a feather flock together.”
- This can help in the identification of individuals with similar characteristics.
  - Marketing campaigns
  - Fraud detection

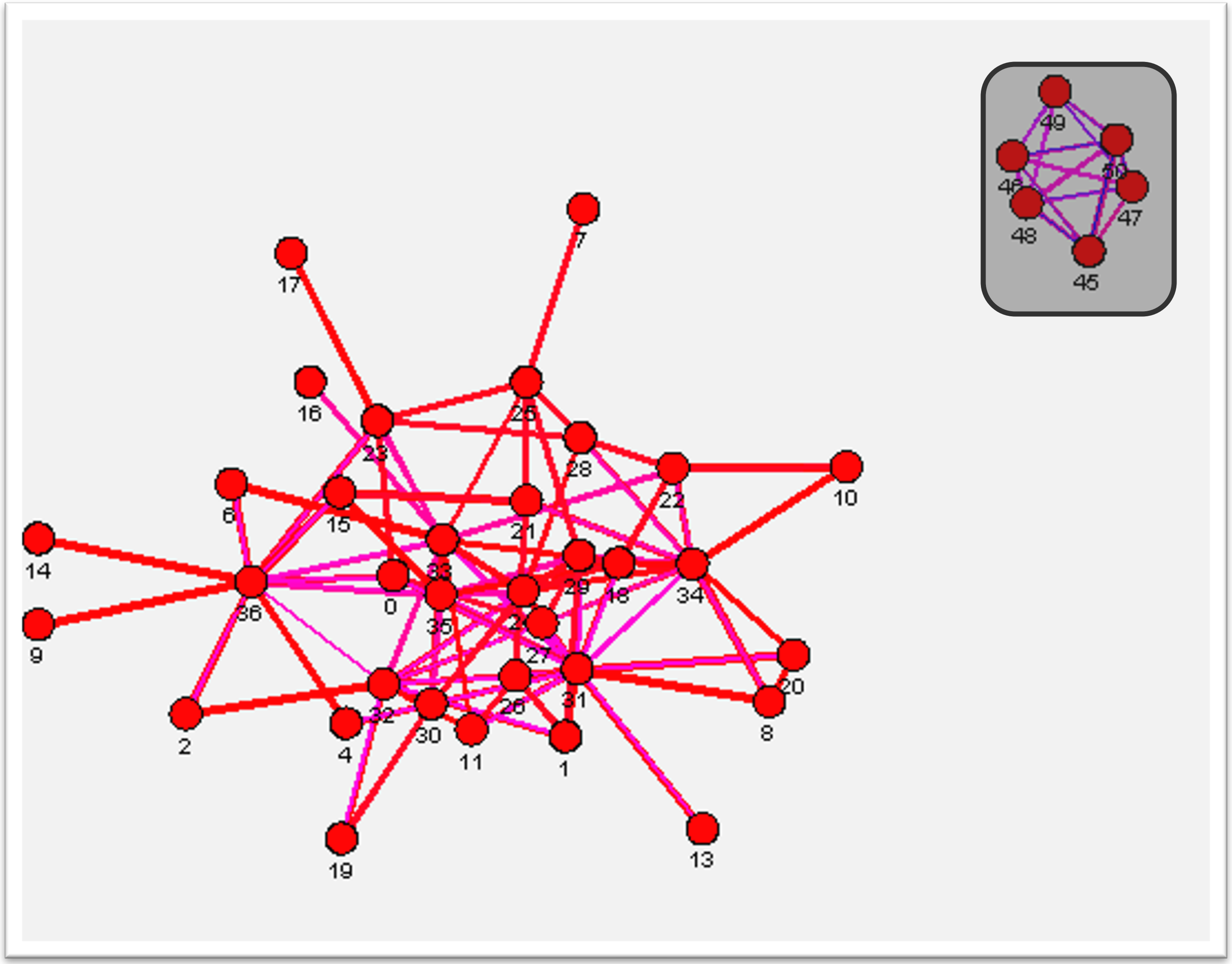
Graph Networks to Check Subgroups



Identify One in a Subgroup



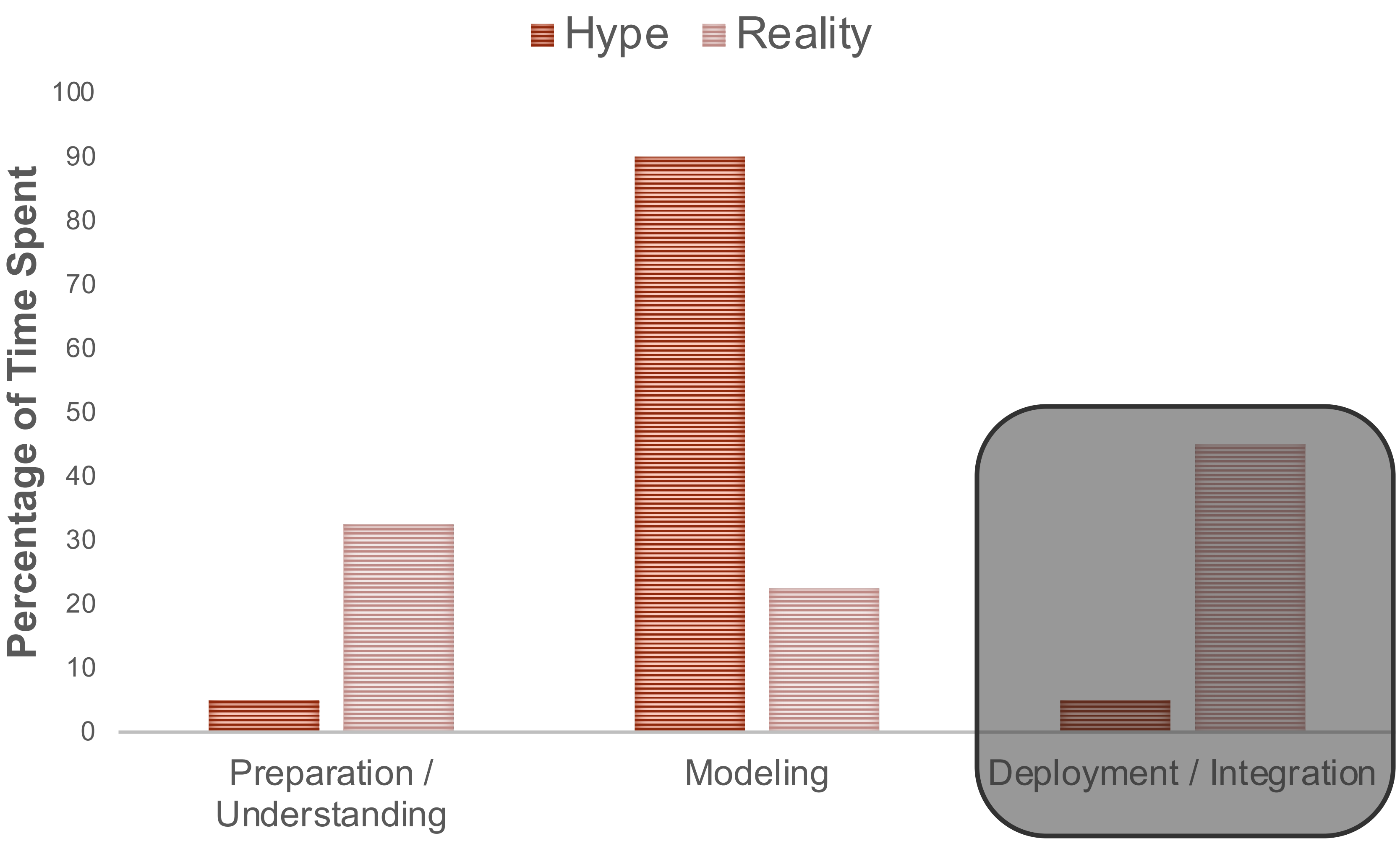
Identify One in a Subgroup → Investigate Subgroup



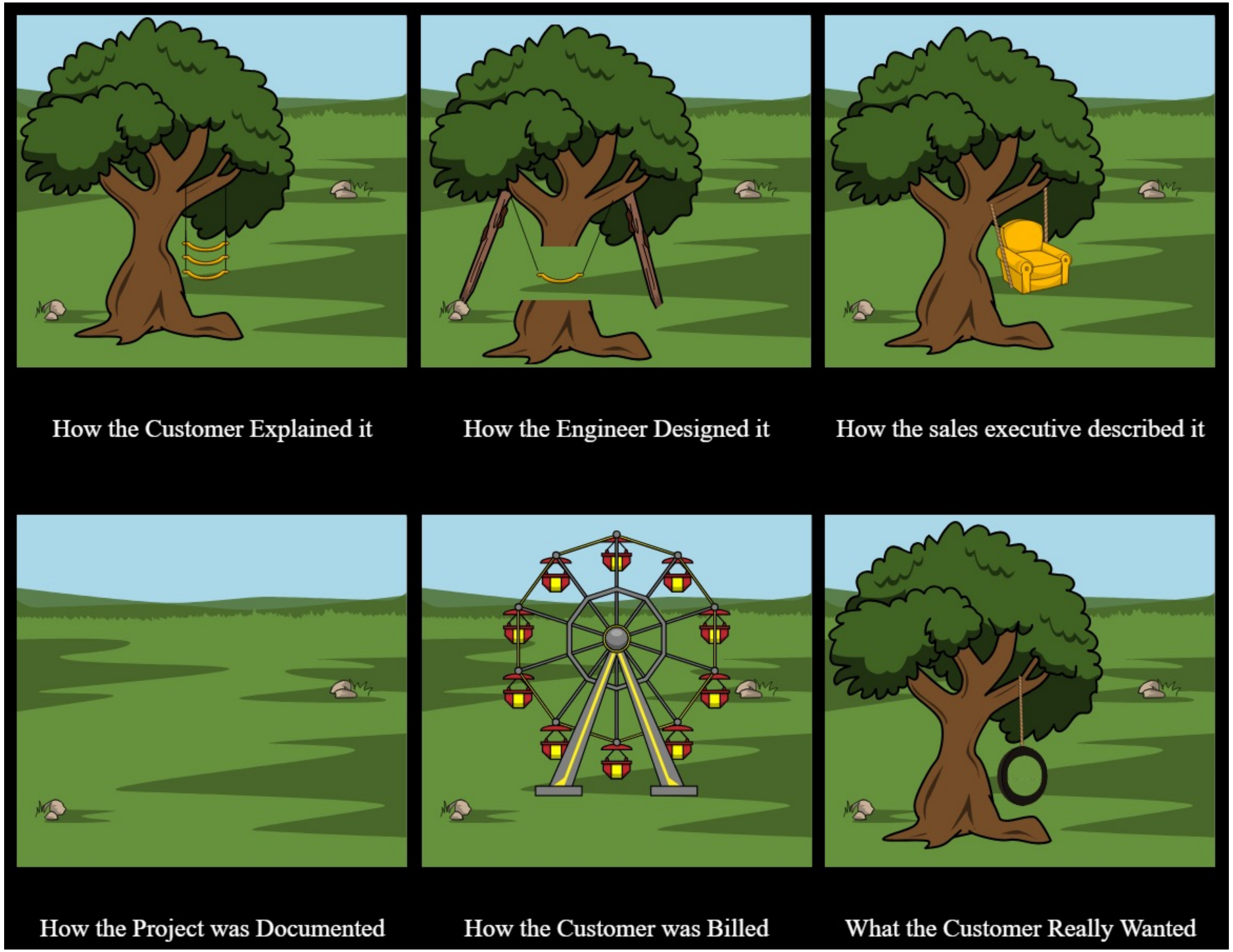
?

INTERPRETABILITY

Data Science Hype vs. Reality



Know Your Customer



Fraud End Users

- Typically, the user of a fraud system is an investigator:
  - Former/current law enforcement
  - Years of experience in investigations
  - Succeeded in their job **without** analytics
  - Have a current process in place
  - Need to be sold on why they might change

Listening

- VERY IMPORTANT
- Listening requires two things:
  1. Desire
  2. Humility

- Research ahead of time – YES!
- Be biased ahead of time – NO!
- Ask many questions to help understand – YES!

Beneficial to Investigators

- Fits into their current process
  - Dashboard?

- Where should I start the investigation?
  - Important variables that drove model to pick this person as potential fraud

Scorecard Models

Variable	Level	Scorecard Points
Pay Time	$x < 10$	100
Pay Time	$10 \leq x < 15$	120
Pay Time	$15 \leq x < 25$	185
Pay Time	$x \geq 25$	200
Report	Yes	225
Report	No	110
Ratio	$x < 1$	225
Ratio	$1 \leq x < 2.5$	200
Ratio	$2.5 \leq x < 5$	180
Ratio	$5 \leq x < 7$	140
Ratio	$x \geq 7$	120

Traffic Light Indicators

Variable	Level	Scorecard Points
Pay Time	$x < 10$	100
Pay Time	$10 \leq x < 15$	120
Pay Time	$15 \leq x < 25$	185
Pay Time	$x \geq 25$	200
Report	Yes	225
Report	No	110
Ratio	$x < 1$	225
Ratio	$1 \leq x < 2.5$	200
Ratio	$2.5 \leq x < 5$	180
Ratio	$5 \leq x < 7$	140
Ratio	$x \geq 7$	120

Traffic Light – Example

Variable	Level	Scorecard Points
Pay Time	$x < 10$	100
Pay Time	$10 \leq x < 15$	120
Pay Time	$15 \leq x < 25$	185
Pay Time	$x \geq 25$	200
Report	Yes	225
Report	No	110
Ratio	$x < 1$	225
Ratio	$1 \leq x < 2.5$	200
Ratio	$2.5 \leq x < 5$	180
Ratio	$5 \leq x < 7$	140
Ratio	$x \geq 7$	120

LONG-TERM FRAUD STRATEGY

Classification

- Claims are referred to the SIU for investigation and classified as fraud or no fraud.
- Investigated claims are labeled “Yes” or “No”.
- Non-investigated claims are labeled “Maybe”.
  - Classified based on unsupervised learning techniques previously discussed.
- All claims are then merged into supervised prediction model.

False Negatives?

- Claims that are labeled as no fraud should occasionally be investigated as well.
- Determine how many low scoring claims can be checked under the budget constraints.
- Randomly select low scoring claims to be passed on to SIU.
- This provides an idea for the false negative rate in the modeling process.

TWO-STAGE FRAUD MODEL

Chance & Loss

- In fraud it is not only important if someone will commit fraud, but how much the fraud will cost the company.
- Want to calculate two things with regards to fraudulent claims:
    1. Probability of fraud occurring
    2. Monetary losses if the fraud occurs

Chance & Loss

- In fraud it is not only important if someone will commit fraud, but how much the fraud will cost the company.
- Want to calculate two things with regards to fraudulent claims:
  1. Probability of fraud occurring
  2. Monetary losses if the fraud occurs

$$Score = P(Fraud) \times E(Loss|Fraud)$$

Chance & Loss

- In fraud it is not only important if someone will commit fraud, but how much the fraud will cost the company.
- Want to calculate two things with regards to fraudulent claims:
  1. Probability of fraud occurring
  2. Monetary losses if the fraud occurs

$$Score = \underbrace{P(Fraud)}_{\text{Binary}} \times \underbrace{E(Loss|Fraud)}_{\text{Continuous}}$$

Chance & Loss

$$Score = P(Fraud) \times E(Loss|Fraud)$$

- There are two typical approaches to handling this type of problem:
  1. Estimate the probability of fraud and the expected loss given fraud as two separate models followed by multiplying them together.
  2. Estimate them jointly in a bivariate model.

Chance & Loss

$$Score = P(Fraud) \times E(Loss|Fraud)$$

- There are two typical approaches to handling this type of problem:
  1. Estimate the probability of fraud and the expected loss given fraud as two separate models followed by multiplying them together.
  2.

Types of Models

- There are some obvious choices for different types of ways to model each of the two models.
- Binary Response Models:
  - Logistic Regression
  - Decision Trees
  - Neural Networks
- Continuous Response Models:
  - Multiple Regression
  - Regression Trees
  - Neural Networks
  - Other

Types of Models

- There are some obvious choices for different types of ways to model each of the two models.
- Binary Response Models:
  - Logistic Regression
  - Decision Trees
  - Neural Networks
- Continuous Response Models:
  - Multiple Regression
  - Regression Trees
  - Neural Networks
  - Other

Types of Models

- What if loss amounts are not available?

- What if there are open claims left in the system?

Types of Models

- What if loss amounts are not available?

- What if there are open claims left in the system?

SURVIVAL ANALYSIS!

- Survival analysis is typically used for fraud modeling to determine the expected loss over time for a claim.
- More common in other types of fraud compared to life insurance.

Chance & Loss

$$Score = P(Fraud) \times E(Loss|Fraud)$$

- There are two typical approaches to handling this type of problem:
  1.
  2. Estimate them jointly in a bivariate model.

Multiple Response Variables

- What happens if you want to model more than one response variable?

Multiple Response Variables

- What happens if you want to model more than one response variable?
  1. Build more than one model

Multiple Response Variables

- What happens if you want to model more than one response variable?
  1. Build more than one model
  2. Multivariate regression models

Multiple Response Variables

- Multivariate regression models model multiple response variables simultaneously.
- Potential to greatly improve accuracy of the models if the response variables are correlated with each other because multivariate models estimate the correlation between them.

Multivariate Regression

- The following is a typical multivariate regression model:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} \beta_{0,1} \\ \beta_{0,2} \\ \vdots \\ \beta_{0,p} \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \cdots & \beta_{1p,1} \\ \vdots & \ddots & \vdots \\ \beta_{p1,1} & \cdots & \beta_{pp,1} \end{bmatrix} \begin{bmatrix} X_{1,1} \\ X_{1,2} \\ \vdots \\ X_{1,p} \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} \beta_{11,k} & \cdots & \beta_{1p,k} \\ \vdots & \ddots & \vdots \\ \beta_{p1,k} & \cdots & \beta_{pp,k} \end{bmatrix} \begin{bmatrix} X_{k,1} \\ X_{k,2} \\ \vdots \\ X_{k,p} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Multivariate Regression

- The following is a typical multivariate regression model:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} \beta_{0,1} \\ \beta_{0,2} \\ \vdots \\ \beta_{0,p} \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \cdots & \beta_{1p,1} \\ \vdots & \ddots & \vdots \\ \beta_{p1,1} & \cdots & \beta_{pp,1} \end{bmatrix} \begin{bmatrix} X_{1,1} \\ X_{1,2} \\ \vdots \\ X_{1,p} \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} \beta_{11,k} & \cdots & \beta_{1p,k} \\ \vdots & \ddots & \vdots \\ \beta_{p1,k} & \cdots & \beta_{pp,k} \end{bmatrix} \begin{bmatrix} X_{k,1} \\ X_{k,2} \\ \vdots \\ X_{k,p} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

$$\boldsymbol{Y} \quad = \quad \boldsymbol{\beta}_0 \quad + \quad \boldsymbol{\beta}_1 \boldsymbol{X}_1 \quad + \cdots + \quad \boldsymbol{\beta}_k \boldsymbol{X}_k \quad + \quad \boldsymbol{\varepsilon}$$

Multivariate Regression

- Let's focus our attention on the bivariate case:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \beta_{0,1} \\ \beta_{0,2} \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \beta_{12,1} \\ \beta_{21,1} & \beta_{22,1} \end{bmatrix} \begin{bmatrix} X_{1,1} \\ X_{1,2} \end{bmatrix} + \cdots + \begin{bmatrix} \beta_{11,k} & \beta_{12,k} \\ \beta_{21,k} & \beta_{22,k} \end{bmatrix} \begin{bmatrix} X_{k,1} \\ X_{k,2} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Bivariate Regression

- Let's focus our attention on the bivariate case:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \beta_{0,1} \\ \beta_{0,2} \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \beta_{12,1} \\ \beta_{21,1} & \beta_{22,1} \end{bmatrix} \begin{bmatrix} X_{1,1} \\ X_{1,2} \end{bmatrix} + \cdots + \begin{bmatrix} \beta_{11,k} & \beta_{12,k} \\ \beta_{21,k} & \beta_{22,k} \end{bmatrix} \begin{bmatrix} X_{k,1} \\ X_{k,2} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

- There are 4 different possibilities for modeling a bivariate case:
  1. Both Y_1 and Y_2 are continuous.
  2. Y_1 is continuous and Y_2 is categorical (binary for now)
  3. Y_2 is continuous and Y_1 is categorical (binary for now)
  4. Both Y_1 and Y_2 are categorical (binary for now).

Thank you!